



IBM SAN Survival Guide

Featuring the Cisco Portfolio

Protect your data with an IBM SAN

Build a SAN too tough to die

Survive and conquer



Jon Tate
Brian Cartwright
John Cronin
Christian Dapprich

ibm.com/redbooks

Redbooks



International Technical Support Organization

**IBM SAN Survival Guide
Featuring the Cisco Portfolio**

October 2003

Note: Before using this information and the product it supports, read the information in “Notices” on page xvii.

First Edition (October 2003)

This edition applies to those products in the IBM/Cisco SAN portfolio.

© Copyright International Business Machines Corporation 2003. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Figures	xi
Tables	xv
Notices	xvii
Trademarks	xviii
Preface	xix
The team that wrote this redbook	xx
Become a published author	xxii
Comments welcome	xxii
Part 1. Survival tactics	1
Chapter 1. Introduction	3
1.1 Beyond disaster recovery	4
1.1.1 Whose responsibility is it?	5
1.1.2 The Internet brings increased risks	6
1.1.3 Planning for business continuity	7
1.2 Using a SAN for business continuance	8
1.2.1 SANs and business continuance	9
1.3 SAN business benefits	10
1.3.1 Storage consolidation and sharing of resources	10
1.3.2 Data sharing	12
1.3.3 Non-disruptive scalability for growth	13
1.3.4 Improved backup and recovery	13
1.3.5 High performance	15
1.3.6 High availability server clustering	15
1.3.7 Improved disaster tolerance	16
1.3.8 Allow selection of “best of breed” storage	16
1.3.9 Ease of data migration	16
1.3.10 Reduced total costs of ownership	17
1.3.11 Storage resources match e-business enterprise needs	17
Chapter 2. SAN fabric components	19
2.1 ASIC technology	20
2.2 Fiber optic interconnects	20
2.2.1 Small Form Factor Optical Transceivers	21
2.2.2 Gigabit Interface Converters	23

2.2.3	Gigabit Link Modules	26
2.2.4	Media Interface Adapters	27
2.2.5	1x9 transceivers	27
2.2.6	Fibre Channel adapter cable	28
2.3	Fibre Channel ports	28
2.3.1	Port types	29
2.4	SAN topologies	30
2.4.1	Point-to-point	31
2.4.2	Arbitrated loop	32
2.4.3	Logins	38
2.4.4	Switched fabric	41
2.4.5	WWN and WWPN	52
2.4.6	Zoning	55
2.4.7	Expanding the fabric	61
2.5	SAN software management standards	63
2.5.1	Application management	64
2.5.2	Data management	65
2.5.3	Resource management	66
2.5.4	Network management	66
2.5.5	Element management	68
2.5.6	Storage Management Initiative	70
2.5.7	InfiniBand	71
2.6	Fabric management methods	72
2.6.1	Common methods	72
2.6.2	Hardware setup for switch management	73
2.6.3	Managing with Telnet	73
2.7	SAN standards	73
2.7.1	SAN industry associations and organizations	74
2.7.2	List of evolved Fibre Channel standards	79
Chapter 3. SAN features		85
3.1	Fabric implementation	86
3.1.1	Blocking	87
3.1.2	Ports	88
3.1.3	Inter-Switch Links	88
3.1.4	RSCN	93
3.2	Classes of service	93
3.2.1	Class 1	94
3.2.2	Class 2	94
3.2.3	Class 3	94
3.2.4	Class 4	95
3.2.5	Class 5	95
3.2.6	Class 6	95

3.2.7 Class F	96
3.2.8 Communication	96
3.2.9 Solutions	96
3.3 Distance	100
3.3.1 Dark fiber	101
3.3.2 Dense Wavelength Division Multiplexing	102
3.3.3 Primary and secondary routes	104
3.4 Time-out values	105
3.4.1 Time-out value settings	105
3.5 Buffers	106
3.6 Data protection	109
3.6.1 RAID	109
3.6.2 Mirroring	110
3.6.3 Clustering	111
3.6.4 Dual pathing	111
3.7 SAN platforms	113
3.7.1 zSeries and S/390	113
3.7.2 pSeries	114
3.7.3 xSeries	114
3.7.4 iSeries	115
3.8 Security	115
3.8.1 Control types	116
3.8.2 Vulnerabilities	116
3.8.3 Fibre Channel security	118
3.8.4 Security mechanisms	119
3.8.5 IP security	121
Chapter 4. SAN disciplines	123
4.1 Floor plan	124
4.1.1 SAN inventory	124
4.1.2 Cable types and cable routing	125
4.1.3 Planning considerations and recommendations	129
4.1.4 Structured cabling	131
4.1.5 Data center fiber cabling options	131
4.1.6 Cabinets	134
4.1.7 Phone sockets	135
4.1.8 Environmental considerations	135
4.1.9 Location	136
4.1.10 Sequence for design	136
4.2 Naming conventions	138
4.2.1 Servers	138
4.2.2 Cabinets	139
4.2.3 SAN fabric components	139

4.2.4 Cables	140
4.2.5 Zones	141
4.3 Documentation	141
4.4 Power-on sequence	142
4.5 Security	142
4.5.1 General	143
4.5.2 Physical access	144
4.5.3 Remote access	144
4.6 Education	145
4.6.1 SAN administrators	146
4.6.2 Skills	146
4.6.3 Certification	146
Chapter 5. Host Bus Adapters	149
5.1 Selection criterion	150
5.1.1 IBM supported HBAs	150
5.1.2 IBM SSG HBA and SAN interoperability matrix	150
5.1.3 ESS host systems attachment	150
5.1.4 Special features	150
5.1.5 Quantity of servers	151
5.1.6 Product specifics	151
5.2 Emulex	151
5.2.1 LP7000E	151
5.2.2 LP8000	152
5.2.3 LP8000S	152
5.2.4 LP9002DC	153
5.2.5 LP9002L	153
5.2.6 LP9002S	153
5.2.7 Emulex special features	154
5.2.8 Device drivers	155
5.2.9 Emulex utilities	155
5.2.10 Installation	157
5.2.11 Management	157
5.2.12 Troubleshooting	158
5.2.13 Performance	160
5.3 JNI	164
5.3.1 FCI-1063-N 32-bit PCI to FC HBA	164
5.3.2 FC64-1063-N 64-bit SBus to FC HBA	164
5.3.3 FCE-1063 64-bit SBus to FC HBA	165
5.3.4 FCE2-1063 64-bit Dual Port SBus to FC HBA	165
5.3.5 FCE-1473 1 or 2 Gb 64-bit SBus to FC HBA	165
5.3.6 FCE-6410 64-bit PCI to FC HBA	165
5.3.7 FCE2-6412 64-bit Dual Port PCI to FC HBA	165

5.3.8	FCC-6460 1 or 2 Gb Compact PCI to FC HBA	166
5.3.9	FCE-6460 1 or 2 Gb 64-bit PCI to FC HBA	166
5.3.10	Drivers	166
5.3.11	Management	167
5.3.12	Installation	168
5.3.13	Troubleshooting the SBUS HBA	169
5.3.14	Troubleshooting the JNI PCI HBA	170
5.3.15	Troubleshooting both PCI and SBUS	170
5.3.16	Performance settings	172
5.4	QLogic	176
5.4.1	QLA2100	176
5.4.2	QLA2200	177
5.4.3	QLA2300	178
5.4.4	Installation	178
5.4.5	Management	178
5.4.6	Troubleshooting	178
5.4.7	Performance	179
Chapter 6.	Cisco directors and switches	185
6.1	Overview	186
6.1.1	Cisco directors and switches	186
6.1.2	Supported attachments	188
6.1.3	Supported port-types	188
6.1.4	Availability	191
6.1.5	Scalable capacity	192
6.2	MDS 9509 Multilayer Director component overview	192
6.2.1	Chassis	192
6.2.2	Backplane	193
6.2.3	Power supply	194
6.2.4	Fan assembly	195
6.2.5	Supervisor module	195
6.3	MDS 9216 Multilayer Fabric Switch	202
6.3.1	Chassis	202
6.3.2	Backplane	203
6.3.3	Power supply	204
6.3.4	Fan assembly	204
6.3.5	Supervisor module	204
6.3.6	Processor	204
6.3.7	Interfaces	205
6.3.8	Switching modules	205
6.4	Cisco SAN capabilities	210
6.4.1	PortChannels	210
6.4.2	Trunking	211

6.4.3 IP services	212
6.4.4 Quality of Service	213
6.5 Software management	215
6.5.1 Cisco MDS 9000 Fabric Manager	215
6.5.2 In-band management and out-of-band management	216
6.5.3 Using the setup routine	218
6.5.4 Controlling administrator access with users and roles	219
6.5.5 Accessing Cisco Fabric Manager	219
6.5.6 Connecting to a supervisor module	220
6.5.7 Software download and install	220
6.5.8 Launching views	221
6.5.9 Discovering and viewing the network fabric	222
6.5.10 Menu bar, toolbars and message bar	225
6.6 Zoning	229
6.6.1 Zone features	230
6.6.2 Zone membership	231
6.6.3 Configuring a zone	231
6.6.4 Zone enforcement	231
6.6.5 Zone sets	232
6.6.6 Default zone	232
6.7 Virtual SAN	233
6.7.1 VSANs versus zones	236
6.7.2 Default and isolated VSANs	236
6.7.3 VSAN membership	237
6.7.4 VSAN attributes	237
6.8 Security management	238
6.8.1 Switch access security	238
6.8.2 User authentication	238
6.8.3 Role-based access	239
6.8.4 User authentication	239
6.8.5 Role-based authorization	240
6.8.6 Accounting	241
6.8.7 Standards	241
6.9 Switch interoperability mode	241
6.10 Troubleshooting	243
6.10.1 Troubleshooting with Fabric Manager	243
6.10.2 Monitoring network traffic using SPAN	245
6.10.3 Monitoring traffic using Fibre Channel analyzers	249
6.11 Interoperability matrix	252
 Chapter 7. General SAN troubleshooting tips	 255
7.1 Overview	256
7.2 Reporting failure	256

7.3 Where to look for failures	258
7.3.1 Connectivity problems	258
7.3.2 Performance problems	260
7.4 Other Fibre Channel diagnostic tools	262
Part 2. Survival solutions	267
Chapter 8. General solutions	269
8.1 Objectives of SAN implementation	270
8.2 Servers and host bus adapters	270
8.2.1 Path and dual-redundant HBA	271
8.2.2 Multiple paths	271
8.3 Software	272
8.4 Storage	272
8.5 Fabric	274
8.5.1 The fabric-is-a-switch approach	274
8.5.2 The fabric-is-a-network approach	276
8.6 High level fabric design	276
8.7 Definitions	280
8.7.1 Port formulas	282
8.8 Our solutions	283
Chapter 9. Cisco director and switch solutions	285
9.1 Performance solutions	286
9.2 Availability solutions	290
9.2.1 Dual fabric	291
9.2.2 Dual fabrics over distance	295
9.3 Distance solutions	300
9.4 Clustering solutions	303
9.5 Secure solutions	308
9.5.1 Zoning security solution	308
9.6 Loop solutions	312
9.6.1 Utilizing the translatative loop port	312
Glossary	317
Related publications	331
IBM Redbooks	331
Other resources	331
Referenced Web sites	332
How to get IBM Redbooks	334
IBM Redbooks collections	334
Index	335

Figures

1-1	Business outage causes	8
1-2	Storage consolidation	11
1-3	Logical storage consolidation	12
1-4	Loading the IP network	14
1-5	SAN total storage solutions	18
2-1	Small Form Factor Transceiver	21
2-2	SFF hot-pluggable transceiver (SFP) with LC connector fiber cable	22
2-3	Dual SC fiber-optic plug connector	25
2-4	Gigabit Interface Converter	25
2-5	Gigabit Link Module	26
2-6	Media Interface Adapter	27
2-7	1x9 transceivers	27
2-8	Fibre Channel adapter cable	28
2-9	Fibre Channel port types	30
2-10	Point-to-point	31
2-11	Arbitrated loop	32
2-12	Private loop implementation	35
2-13	Public loop implementation	36
2-14	Fibre Channel logins	39
2-15	Sample switched fabric configuration	41
2-16	Fabric port address	43
2-17	Arbitrated loop address translation	46
2-18	Meshed topology switched fabric	49
2-19	Fabric shortest path first	51
2-20	World Wide Name addressing scheme	53
2-21	WWN and WWPN	54
2-22	WWN and WWPN entries in a name server table	55
2-23	Zoning	56
2-24	Zoning based on the switch port-number	57
2-25	Zoning based on the devices WWN	59
2-26	Cascading in a switched fabric	62
2-27	SAN management hierarchy	64
2-28	Common Interface Model for SAN management	65
2-29	Typical SAN environment	67
2-30	Device management elements	68
2-31	Groups involved in setting storage management standards	74
3-1	Cascading directors	86
3-2	Non-blocking and blocking switching	87

3-3	Parallel ISLs - low traffic	89
3-4	Parallel ISLs - high traffic	89
3-5	ISL Trunking.	90
3-6	Four-switch fabric.	91
3-7	Channel distances comparison	100
3-8	Distance solution	101
3-9	Distance solution with Finisar extenders	102
3-10	DWDM point-to-point configuration	103
3-11	DWDM hubbed ring configuration	104
3-12	Adjacent FC devices	107
3-13	Vulnerabilities.	117
4-1	Mode differences through the fiber-optic cable	126
4-2	Messy cabling, no cabinet, and no cable labels.	135
4-3	Naming convention development.	140
5-1	Emulex LP8000 HBA	152
5-2	Emulex 9002S HBA (SBUS)	154
5-3	Emulex — port driver screen	156
5-4	Emulex adapter settings	160
5-5	Picture of JNI's FC64_1063 HBA.	164
5-6	EZ Fibre configuration panel	168
5-7	The FCI-1063_N 64 Optical device	169
5-8	QLogic HBA card	177
6-1	MDS 9509 Multilayer Director (IBM 2062-D07)	187
6-2	MDS 9216 Multilayer Switch (IBM 2062-D01)	188
6-3	Cisco MDS 9000 family port types.	191
6-4	Cisco MDS 9509 Multilayer Director layout	193
6-5	16 port switching module	199
6-6	32 port switching module	200
6-7	Cisco MDS 9216 Multilayer Fabric Switch layout	203
6-8	16 port switching module	207
6-9	32 port switching module	208
6-10	PortChannels and ISLs on the Cisco MDS 9000 family of switches . .	211
6-11	Trunking and PortChanneling	212
6-12	Forward Congestion Control	214
6-13	Out-of-band management connection	217
6-14	In-band management connection	218
6-15	Cisco MDS 9000 family - Supervisor Web page	220
6-16	Java Web start warning	221
6-17	<i>Fabric Manager - login screen</i>	222
6-18	Cisco MDS 9000 Fabric Manager - discovery process	223
6-19	Fabric Manager - fabric view	224
6-20	Fabric Manager - VSAN/Switches pane	226
6-21	Fabric Manager - Device View.	228

6-22	Fabric Manager - Summary View	229
6-23	Zoning overview	230
6-24	Traditional SAN	234
6-25	Cisco Virtual SAN	235
6-26	Security with local authentication	239
6-27	Security with RADIUS server	240
6-28	SPAN destination ports	245
6-29	SD_Port for incoming traffic (ingress direction)	246
6-30	SD_Port for outgoing traffic (egress direction)	246
6-31	Fibre Channel analyzer without SPAN	250
6-32	Fibre Channel analyzer using SPAN	251
6-33	Using a single SD_Port to monitor traffic	252
7-1	A simple fabric	257
7-2	AIX error log	258
7-3	pSeries HBA status	259
7-4	SDD information	259
7-5	ESS error log	260
7-6	Bit error formula	261
7-7	FINISAR GTX TraceView	264
7-8	Xyratex FCI Protocol Analyzer	265
8-1	Fabric in a director	275
8-2	Two examples of switch cascading	277
8-3	Ring design	277
8-4	Meshed network design	278
8-5	Host-tier and storage-tier	279
8-6	Tier to tier	279
8-7	Core-edge design	280
9-1	High performance design	286
9-2	Expanding the SAN fabric via E_Ports	289
9-3	Traditional dual fabric design (without VSANs)	291
9-4	Dual fabric design with VSANs	292
9-5	Traditional dual fabric design over distance	295
9-6	Dual fabric design over distance using VSANs	296
9-7	Solution using long distance SFPs	300
9-8	Clustered solution using Cisco MDS 9509 Director	303
9-9	Protecting your data from both human error and sabotage	310
9-10	Utilizing the Cisco MDS 9000 TL_Port	313

Tables

2-1	Distances using SFF-based fiber optics	22
2-2	Distance using 1 Gb/s GBIC based fiber optics	24
2-3	Distance using 2 Gb/s GBIC based fiber-optics	24
2-4	Name server entries	47
2-5	Phantom addresses	47
2-6	Comparison of management access methods	72
2-7	T11 projects	79
3-1	RAID levels definitions	110
4-1	Comparison between structured and non-structured cables	129
4-2	Checklist for proposed layout	137
5-1	LED HBA status	159
5-2	LP8000 adapter recommended configuration file parameters	161
5-3	LP9000 adapter recommended configuration file parameters	162
5-4	Emulex firmware structure	163
5-5	IBM recommended settings for JNI FC64-1063 and JNI FCI-1063	173
5-6	JNI miscellaneous recommended settings	174
5-7	QLA2200F recommendations	181
5-8	QLogic QLA2310F, QLA2340, QLA2342 recommendations	182
6-1	Power supply voltages	194
6-2	Power consumption by module	194
6-3	Processor memory options	196
6-4	Distance specifications for SFP optics	201
6-5	Physical specifications for the MDS 9509	202
6-6	Distance specifications for SFP optics	209
6-7	Physical specifications for the MDS 9216	210
6-8	VSAN versus zoning	236
6-9	Interoperability mode changes	242
6-10	Cisco MDS 9509 Multilayer Director interoperability matrix	253
6-11	Cisco MDS 9216 Multilayer Switch interoperability matrix	254

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.


This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

AFS®	Lotus®	Storage Tank™
AIX®	Magstar®	StorWatch™
AS/400®	Netfinity®	System/38™
DB2®	Notes®	Tivoli®
Enterprise Storage Server®	NUMA-Q®	TotalStorage®
@server™	OS/390®	Wave®
ESCON®	Parallel Sysplex®	WebSphere®
FICON™	pSeries®	xSeries®
FlashCopy®	Redbooks™	z/OS®
IBM®	Redbooks (logo)  ™	zSeries®
ibm.com®	S/390®	
iSeries™	SANergy™	

The following terms are trademarks of other companies:

Intel, Intel Inside (logos), and Pentium are trademarks of Intel Corporation in the United States, other countries, or both.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States, other countries, or both.

UNIX is a registered trademark of The Open Group in the United States and other countries.

Other company, product, and service names may be trademarks or service marks of others.

Preface

As we all know, large ocean going ships never collide with icebergs. If this is the case, there would not seem to be much of a market opening for this IBM® Redbook. However, occasionally life deals out some unexpected pleasures for us to cope with. Surviving any disaster in life is usually a lot easier if you have prepared adequately by taking into account the likely problems, solutions, and their implementation. This can be summarized by the rule of the six P's:

Proper Preparation Prevents Pretty Poor Performance

In this IBM Redbook, we limit ourselves to those situations in which it is likely that a SAN will be deployed. We discuss considerations that need to be taken into account to ensure that no SAN designer is hoisted by their own petard. We present the IBM SAN portfolio of Cisco products, going a little under the surface to show the fault tolerant features that they utilize, and then describe solutions we have designed with all these features taken into account.

This book features the Cisco portfolio. The MDS 9000 family introduces SAN capabilities, including Cisco's Virtual SAN (VSAN) technology, designed to enable efficient SAN utilization by creating hardware-based isolated environments within a single SAN fabric. Each VSAN can be zoned as a typical SAN and maintains its own fabric services for added scalability and resilience.

Each of the solutions that we show has been built on the blood, sweat, and tears of practical experience in the House of SAN. The designs themselves have been built, in some cases with cost in mind, in some cases with no cost in mind.

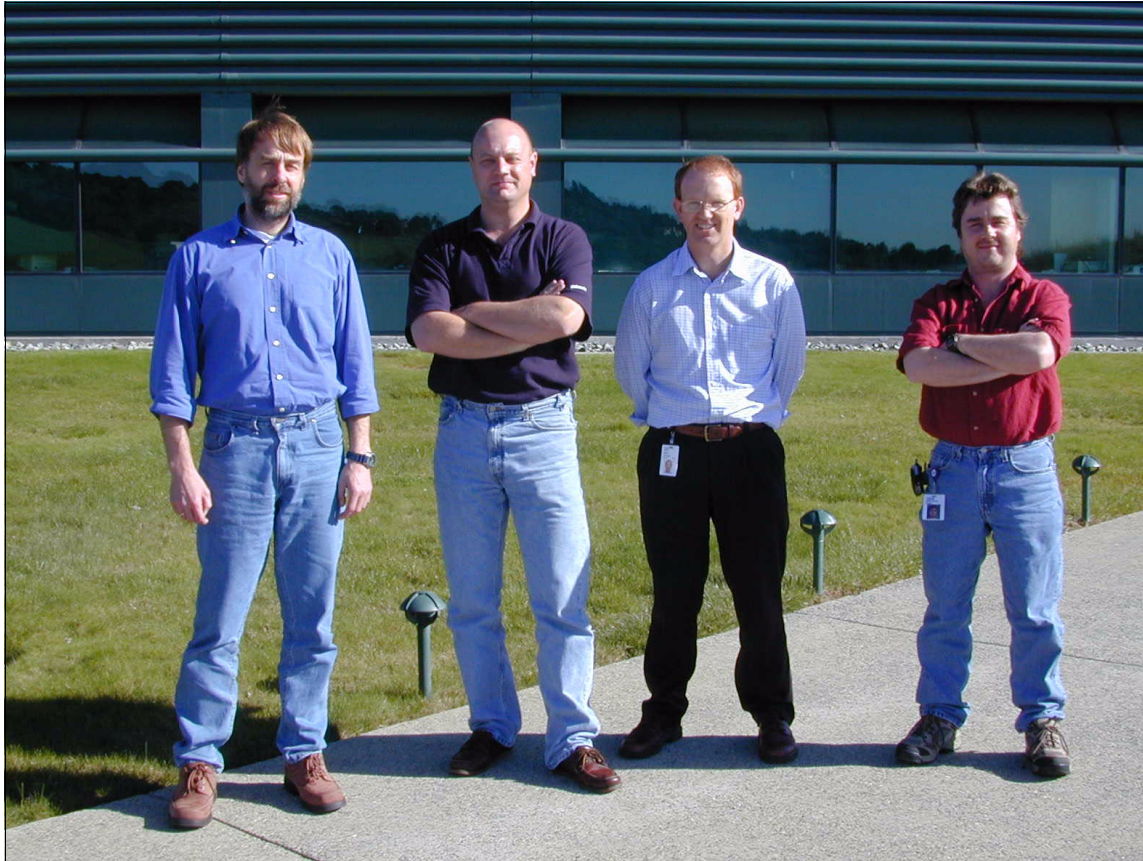
We understand that in this financially constrained world, life is not so easy as we describe it on paper. But that does not make the solutions any less valid. Any good, well thought out SAN design will have taken each concern into account, and either formulated a solution for it, or ignored it, but nonetheless understanding the potential exposure.

IBM brings a vast amount of muscle to the SAN arena, and in this redbook, we have two objectives. First, we position the IBM SAN products that are currently in our portfolio. Second, we show how those products can be configured together to build a SAN that not only allows you to survive most forms of disaster, but also provides performance benefits.

So, make sure that you know what to do if you hit an iceberg!

The team that wrote this redbook

This redbook was produced by a team of specialists from Australia, Germany, the US, and the UK working at the International Technical Support Organization, San Jose Center.



L-R: Christian, Jon, Brian, and John

Jon Tate is a Project Manager for IBM TotalStorage® SAN Solutions at the International Technical Support Organization, San Jose Center. Before joining the ITSO in 1999, he worked in the IBM Technical Support Center, providing Level 2 support for IBM storage products. Jon has 17 years of experience in storage software and management, services, and support, and is both an IBM Certified IT Specialist, and an IBM SAN Certified Specialist.

Brian Cartwright is a Senior Storage Specialist based in Brisbane, Australia. He has 16 years of experience in the IT industry, with 12 years being spent in the storage area, including 7 years in IBM. He holds a degree in Computing Studies from Canberra University. His areas of expertise include Storage Area Networks, Disk and Tape subsystems, and IBM storage software. Brian has written extensively on IBM SAN solutions, and is an IBM SAN, Disk, and Tape Certified Specialist.

John Cronin is a Systems Management Integration Professional for IBM Global Services in Phoenix, Arizona. He has 14 years of experience in IT. He spent the first 12 years in Mainframe with eight years in Storage Management. Since joining IBM two years ago, John has specialized in SAN Management and troubleshooting. His areas of expertise include SMS/HSM, Local Area Networking, and Storage Area Networking. He is a Cisco Certified Network Professional as well as a SNIA Certified Fibre Channel Professional.

Christian Dapprich is an ITS Network Specialist in IBM EMEA SAN Central Support Center in Germany. He has 25 years of experience in IT and Telecommunications, including 12 years with IBM, and holds a degree in Electrical Engineering from the FH des Saarlandes. His areas of expertise include Local Area, Wide Area, and Storage Area Networking, and he is an IBM SAN Certified Specialist. He has written extensively on IBM SAN switches and troubleshooting.

Thanks to the following people for their contributions to this project:

Omar Escola
Nick Milsom
Kum Wai Wong
The previous authors of this redbook

Tom Cady
Yvonne Lyon
Deanna Polm
Emma Jacobs
Sokkieng Wang
International Technical Support Organization, San Jose Center

Edith Kropf
Peter Thurston
Diana Tseng
Karen Ward
Michelle Wright
Ruoyi Zhou
IBM Storage Systems Group

Dan Hersey
Cisco Systems

Richio Aikawa
Jon Krueger
Emulex Corporation

Charles Portnoy
JNI Corporation

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions, while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you'll develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our Redbooks™ to be as helpful as possible. Send us your comments about this or other Redbooks in one of the following ways:

- Use the online **Contact us** review redbook form found at:

ibm.com/redbooks

- Send your comments in an Internet note to:

redbook@us.ibm.com

- Mail your comments to:

IBM Corporation, International Technical Support Organization
Dept. QXXE Building 80-E2
650 Harry Road
San Jose, California 95120-6099



Part 1

Survival tactics

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”

Sherlock Holmes.

As we have no wish to make a capital mistake, we will describe some of the components, features, and disciplines that are an essential piece of the survival jigsaw.



Introduction

Until recently, disaster planning for businesses focused on recovering centralized data centers following a catastrophe, either natural or man-made. While these measures remain important to disaster planning, the protection they provide is far from adequate for today's distributed computing environments.

The goal for companies today is to achieve a state of business continuity, where critical systems and networks are always available. To attain and sustain business continuity, companies must engineer availability, security, and reliability into every process from the outset.

In this chapter we consider the many benefits SAN has to offer in these areas.

1.1 Beyond disaster recovery

When disaster recovery emerged as a formal discipline and a commercial business in the 1980s, the focus was on protecting the data center — the heart of a company's heavily centralized IT structure. This model began to shift in the early 1990s to distributed computing and client/server technology.

At the same time, information technology became embedded in the fabric of virtually every aspect of a business. Computing was no longer something done in the background. Instead, critical business data could be found across the enterprise — on desktop PCs and departmental local area networks, as well as in the data center.

This evolution continues today. Key business initiatives such as enterprise resource planning (ERP), supply chain management, customer relationship management, and e-business have all made continuous, ubiquitous access to information crucial to an organization. This means business can no longer function without information technology in the following areas:

- ▶ Data
- ▶ Software
- ▶ Hardware
- ▶ Networks
- ▶ Call centers
- ▶ Laptop computers

A company that sells products on the Web, for example, or supports customers with an around-the-clock call center, must be operational 24 hours a day, 7 days a week, or customers will go elsewhere. An enterprise that uses e-business to acquire and distribute parts and products is not only dependent on its own technology but that of its suppliers. As a result, protecting critical business processes, with all their complex interdependencies, has become as important as safeguarding data itself.

The goal for companies with no business tolerance for downtime is to achieve a state of business continuity, where critical systems and networks are continuously available, no matter what happens. This means thinking proactively: engineering availability, security, and reliability into business processes from the outset — not retrofitting a disaster recovery plan to accommodate ongoing business continuity requirements.

1.1.1 Whose responsibility is it?

Many senior executives and business managers consider business continuity the responsibility of the IT department. However, it is no longer sufficient or practical to vest the responsibility exclusively in one group. Web-based and distributed computing have made business processes too complex and decentralized. What's more, a company's reputation, customer base and, of course, revenue and profits are at stake. All executives, managers, and employees must therefore participate in the development, implementation, and ongoing support of continuity assessment and planning.

The same information technology driving new sources of competitive advantage has also created new expectations and vulnerabilities. On the Web, companies have the potential to deliver immediate satisfaction — or dissatisfaction — to millions of people. Within ERP and supply chain environments, organizations can reap the rewards of improved efficiencies, or feel the impact of a disruption anywhere within their integrated processes.

With serious business interruption now measured in minutes rather than hours, even success can bring about a business disaster. Web companies today worry more about their ability to handle unexpected peaks in customer traffic than about fires or floods — and for good reason. For example, an infrastructure that cannot accommodate a sudden 200 percent increase in Web site traffic generated by a successful advertising campaign can result in missed opportunities, reduced revenues, and a tarnished brand image. Because electronic transactions and communications take place so quickly, the amount of work and business lost in an hour far exceeds the toll of previous decades. According to reports, the financial impact of a major system outage can be enormous:

- ▶ US\$6.5 million per hour in the case of a brokerage operation
- ▶ US\$2.6 million per hour for a credit-card sales authorization system
- ▶ US\$14,500 per hour in automated teller machine (ATM) fees if an ATM system is offline

Even what was once considered a minor problem — a faulty hard drive or a software glitch — can cause the same level of loss as a power outage or a flooded data center, if a critical business process is affected. For example, it has been calculated that the average financial loss per hour of disk array downtime stands at:

- ▶ US\$29,301 in the securities industry
- ▶ US\$26,761 for manufacturing
- ▶ US\$17,093 for banking
- ▶ US\$9,435 for transportation

More difficult to calculate are the intangible damages a company can suffer: lower morale and productivity, increased employee stress, delays in key project time lines, diverted resources, regulatory scrutiny, and a tainted public image. In this climate, executives responsible for company performance now find their personal reputations at risk. Routinely, companies that suffer online business disruptions for any reason make headlines the next day, with individuals singled out by the press. Moreover, corporate directors and officers can be liable for the consequences of business interruption or loss of business-critical information. Most large companies stipulate in their contracts that suppliers must deliver services or products under any circumstances. What's more, adequate protection of data may be required by law, particularly for a public company, financial institution, utility, health care organization, or government agency.

Together, these factors make business continuity the shared responsibility of an organization's entire senior management, from the CEO to line-of-business executives in charge of crucial business processes. Although IT remains central to the business continuity formula, IT management alone cannot determine which processes are critical to the business and how much the company should pay to protect those resources.

1.1.2 The Internet brings increased risks

A recent IBM survey of 226 business recovery corporate managers revealed that only eight percent of Internet businesses are prepared for a computer system disaster. Yet doing business online means exposing many business-critical applications to a host of new risks. While the Internet creates tremendous opportunity for competitive advantage, it can also give partners, suppliers, customers, employees and hackers increased access to corporate IT infrastructures. Unintentional or malicious acts can result in a major IT disruption. Moreover, operating a Web site generates organizational and system-related interdependencies that fall outside of a company's control from Internet Service Providers (ISP) and telecommunications carriers to the hundreds of millions of public network users.

Therefore, the greatest risk to a company's IT operations may no longer be a hurricane, a 100-year flood, a power outage, or even a burst pipe. Planning for continuity in an e-business environment must address vulnerability to network attacks, hacker intrusions, viruses, and spam, as well as ISP and telecommunication line failures.

1.1.3 Planning for business continuity

Few organizations have the need or the resources to assure business continuity equally for every functional area. Therefore, any company that has implemented a single business continuity strategy for the entire organization is likely under-prepared, or spending money unnecessarily. The key to business continuity lies in understanding your business, determining which processes are critical to staying in that business, and identifying all the elements crucial to those processes. Specialized skills and knowledge, physical facilities, training, and employee satisfaction, as well as information technology, should all be considered. By thoroughly analyzing these elements, you can accurately identify potential risks and make informed business decisions about accepting, mitigating or transferring those risks.

Once you have developed a program for assuring that critical processes will be available around the clock, you should assume that it will fail — and commit to keeping your program current with business and technology infrastructure changes. A fail-safe strategy assumes that no business continuity program can provide absolute protection from every type of damage — no matter how comprehensive your high-availability, redundancy, fault tolerance, clustering, and mirroring strategies.

Today, the disasters most likely to bring your business to a halt are the result of human error or malice: the employee who accidentally deletes a crucial block of data; the disgruntled ex-employee seeking revenge by introducing a debilitating virus; the thief who steals vital trade secrets from your mainframe; or the hacker who invades your network. According to a joint study by the U.S. Federal Bureau of Investigation and the Computer Security Institute, the number and severity of successful corporate hacks is increasing dramatically, particularly intrusions by company insiders. In one study, 250 Fortune 1000 companies reported losses totaling US\$137 million in 1997 — an increase of 37 percent over the previous year.

Making an executive commitment to regularly testing, validating, and refreshing your business continuity program can protect your company against perhaps the greatest risk of all — complacency. In the current environment of rapid business and technology change, even the smallest alteration to a critical application or system within your enterprise or supply chain can cause an unanticipated failure, impacting your business continuity. Effective business protection planning addresses not only what you need today, but what you will need tomorrow and into the future.

1.2 Using a SAN for business continuance

Although some of the concepts that we detail purely apply to only the SAN environment, there are general considerations that need to be taken into account in any environment. Any company that is serious about business continuance will have considered and applied processes or procedures to take into account any of the eventualities that may occur, such as those listed in Figure 1-1.

A/C Failure	Evacuation	Low Voltage	Sprinkler
Acid Leak	Explosion	Microwave Fade	Discharge
Asbestos	Fire	Network Failure	Static
Bomb Threat	Flood	PCB	Electricity
Bomb Blast	Fraud	Contamination	Strike Action
Brown Out	Frozen Pipes	Plane Crash	S/W Error
Burst Pipe	Hacker	Power Outage	S/W Ransom
Cable Cut	Hail Storm	Power Spike	Terrorism
Chemical Spill	Halon	Power Surge	Theft
CO Fire	Discharge	Programmer	Toilet Overflow
Condensation	Human Error	Error	Tornado
Construction	Humidity	Raw Sewage	Train
Coolant Leak	Hurricane	Relocation Delay	Derailment
Cooling Tower	HVAC Failure	Rodents	Transformer
Leak	H/W Error	Roof Cave In	File
Corrupted Data	Ice Storm	Sabotage	Tsunami
Diesel	Insects	Shotgun Blast	UPS Failure
Generator	Lightening	Shredded Data	Vandalism
Earthquake	Logic Bomb	Sick building	Vehicle Crash
Electrical Short	Lost Data	Smoke Damage	Virus
Epidemic		Snow Storm	Water (Various)
			Wind Storm
			Volcano

Figure 1-1 Business outage causes

Some of these problems are not necessarily common to all regions throughout the world, but they should be considered nonetheless, even if only to dismiss the eventuality that they may happen. Careful consideration will result in a deeper understanding of what is likely to cause a business outage, rather than just adopting an attitude that says “it will not happen to me”. After all, remember that the Titanic was unsinkable.

1.2.1 SANs and business continuance

So why would the risk increase if you were to implement a SAN in your environment? The short answer is that it may not increase the risk. It may expose you to more risk over a greater area, for example, the SCSI 25 m restriction means that a small bomb planted in the correct position would do quite nicely. If you are using a SAN for distance solutions, then it might be necessary to increase the size of the bomb, or plant many more of them, to cause the same effect.

What a SAN means is that you now are beginning to explore the potential for ensuring that your business can actually continue in the wake of a disaster. It may be able to do this by:

- ▶ Providing for greater operational distances
- ▶ Providing mirrored storage solutions for local disasters
- ▶ Providing failover support for local disasters
- ▶ Providing remote vaulting anywhere in the world
- ▶ Providing high availability file serving functionality
- ▶ Providing the ability to avoid space outage situations for higher availability

If we are to take the simple example of distance, what a SAN will allow you to do is to break the SCSI distance barrier. Does this in itself make you any safer? Of course it doesn't. Does it give you an opportunity to minimize the risk to your business. Of course it does.

It is up to you if you decide to use that to your advantage, or ignore it and the other benefits that it may bring to your business. One thing is certain though; if you don't exploit the SAN's potential to its fullest, other people may. Those other people might be your competitors. Does that worry you? If it doesn't, then you may want to stop reading right now, because this redbook is not for you! We are targeting those people that are concerned with unleashing the potential of their SAN, or are interested in seeing what a SAN can do.

But that is not all we will do. We will provide you with as much information as we can, that will cover the data center environment from floor to ceiling and the considerations that you should take to ensure minimal exposure to any outage.

As availability is linked to business continuance and recovery, we will also cover methods that can be employed to ensure that the data in your SAN is available to those that are authorized to access it, and protected from those that aren't.

1.3 SAN business benefits

Today's business environment creates many challenges for the enterprise IT planner. This is a true statement and relates to more than just business continuance, so perhaps now is a good time to look at whether deploying a SAN will solve more than just one problem. It may be an opportunity to look at where you are today and where you want to be in three years' time. Is it better to plan for migration to a SAN from the start, or try to implement one later after other solutions have been considered and possibly implemented? Are you sure that the equipment that you install today will still be usable three years later? Is there any use that you can make of it outside of business continuance? A journey of a thousand miles begins with one step.

In the topics that follow we will remind you of some of the business benefits that SANs can provide. We have identified some of the operational problems that a business faces today, and which could potentially be solved by a SAN implementation.

1.3.1 Storage consolidation and sharing of resources

By enabling storage capacity to be connected to servers at a greater distance, and by disconnecting storage resource management from individual hosts, a SAN enables disk storage capacity to be consolidated. The results can be lower overall costs through better utilization of the storage, lower management costs, increased flexibility, and increased control.

This can be achieved physically or logically, as we explain in the following sections.

Physical consolidation

Data from disparate storage subsystems can be combined on to large, enterprise class shared disk arrays, which may be located at some distance from the servers. The capacity of these disk arrays can be shared by multiple servers, and users may also benefit from the advanced functions typically offered with such subsystems. This may include RAID capabilities, remote mirroring, and instantaneous data replication functions, which might not be available with smaller, integrated disks. The array capacity may be partitioned, so that each server has an appropriate portion of the available gigabytes.

Physical consolidation of storage is shown in Figure 1-2.

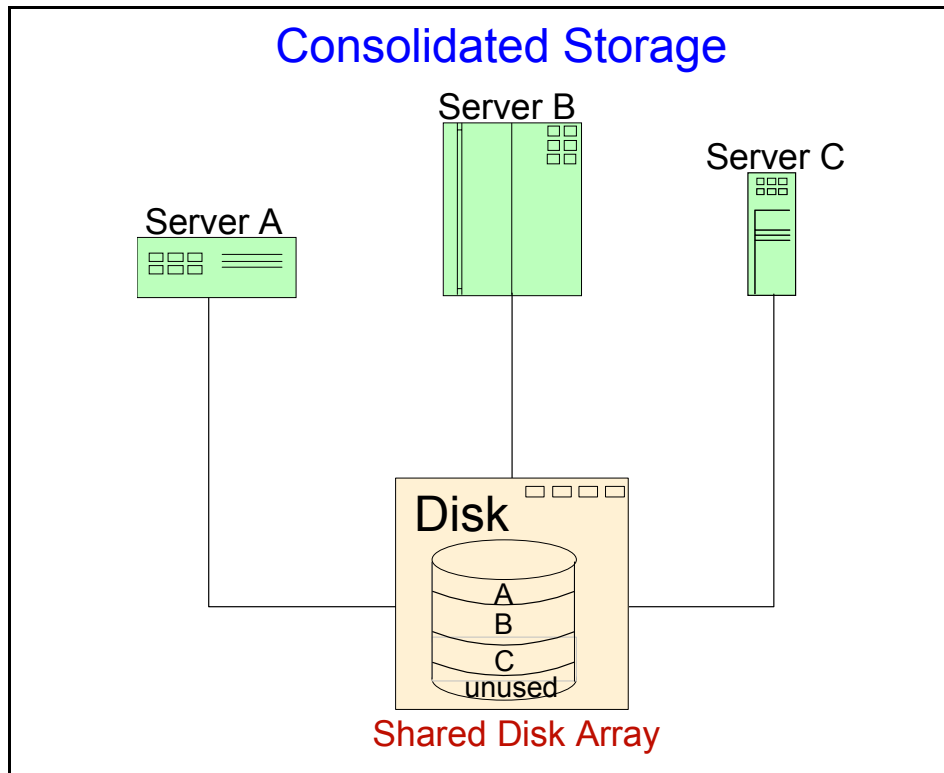


Figure 1-2 Storage consolidation

Available capacity can be dynamically allocated to any server requiring additional space. Capacity not required by a server application can be re-allocated to other servers. This avoids the inefficiency associated with free disk capacity attached to one server not being usable by other servers. Extra capacity may be added, in a non-disruptive manner.

Logical consolidation

It is possible to achieve shared resource benefits from the SAN, but without moving existing equipment. A SAN relationship can be established between a client and a group of storage devices that are not physically co-located (excluding devices which are internally attached to servers). A logical view of the combined disk resources may allow available capacity to be allocated and re-allocated between different applications running on distributed servers, to achieve better utilization. Consolidation is covered in greater depth in *IBM Storage Solutions for Server Consolidation*, SG24-5355.

In Figure 1-3 we show a logical consolidation of storage.

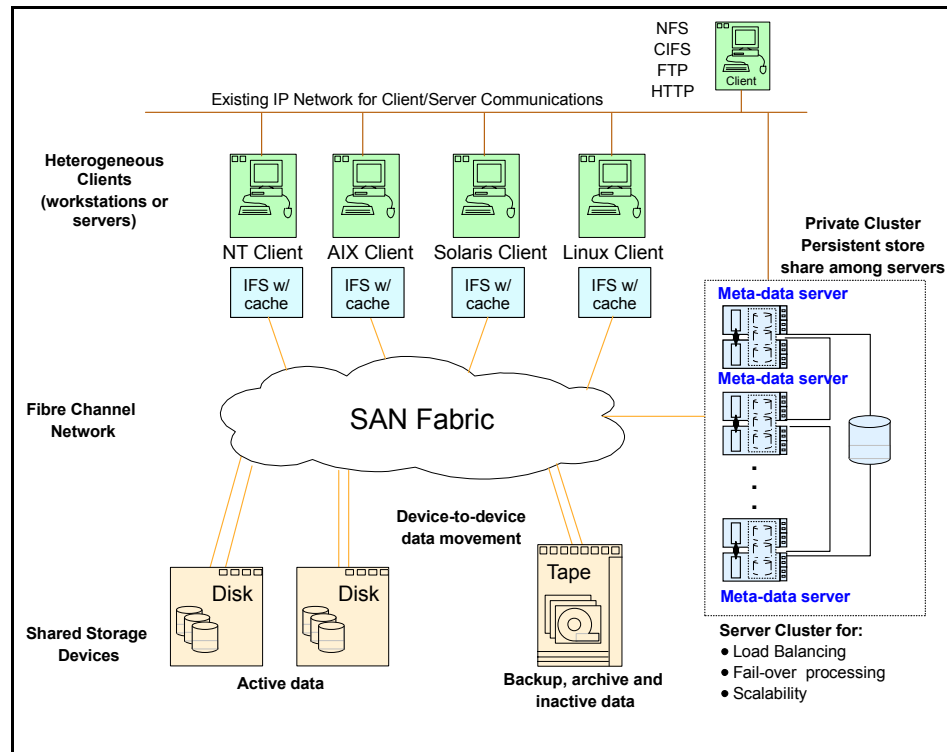


Figure 1-3 Logical storage consolidation

1.3.2 Data sharing

The term “data sharing” is used somewhat loosely by users and some vendors. It is sometimes interpreted to mean the replication of files or databases to enable two or more users, or applications, to concurrently use separate copies of the data. The applications concerned may operate on different host platforms. A SAN may ease the creation of such duplicated copies of data using facilities such as remote mirroring.

Data sharing may also be used to describe multiple users accessing a single copy of a file. This could be called “true data sharing”. In a homogeneous server environment, with appropriate application software controls, multiple servers may access a single copy of data stored on a consolidated storage subsystem.

If attached servers are heterogeneous platforms (for example, with a mix of UNIX and Windows® NT), sharing of data between such unlike operating system environments is complex. This is due to differences in file systems, data formats,

and encoding structures. IBM, however, uniquely offers a true data sharing capability, with concurrent update, for selected heterogeneous server environments, using the Tivoli® SANergy™ File Sharing solution.

The SAN advantage in enabling enhanced data sharing may reduce the need to hold multiple copies of the same file or database. This reduces duplication of hardware costs to store such copies. It also enhances the ability to implement cross enterprise applications, such as e-business, which may be inhibited when multiple data copies are stored.

1.3.3 Non-disruptive scalability for growth

There is an explosion in the quantity of data stored by the majority of organizations. This is fueled by the implementation of applications, such as e-business, e-mail, Business Intelligence, Data Warehouse, and Enterprise Resource Planning. Some industry analysts, such as IDC and Gartner Group, estimate that electronically stored data is doubling every year. In the case of e-business applications, opening the business to the Internet, there have been reports of data growing by more than 10 times annually. This is a nightmare for planners, as it is increasingly difficult to predict storage requirements.

A finite amount of disk storage can be connected physically to an individual server due to adapter, cabling, and distance limitations. With a SAN, new capacity can be added as required, without disrupting ongoing operations. SANs enable disk storage to be scaled independently of servers.

1.3.4 Improved backup and recovery

With data doubling every year, what effect does this have on the backup window? Backup to tape, and recovery, are operations which are problematic in the parallel SCSI or LAN based environments. For disk subsystems attached to specific servers, two options exist for tape backup. Either it must be done onto a server attached tape subsystem, or by moving data across the LAN.

Tape pooling

Providing tape drives to each server is costly, and also involves the added administrative overhead of scheduling the tasks, and managing the tape media. SANs allow for greater connectivity of tape drives and tape libraries, especially at greater distances. Tape pooling is the ability for more than one server to logically share tape drives within an automated library. This can be achieved by software management, using tools, such as Tivoli Storage Manager; or with tape libraries with outboard management, such as the IBM 3494.

LAN-free and server-free data movement

Backup using the LAN moves the administration to centralized tape drives or automated tape libraries. However, at the same time, the LAN experiences very high traffic volume during the backup or recovery operations, and this can be extremely disruptive to normal application access to the network. Although backups can be scheduled during non-peak periods, this may not allow sufficient time. Also, it may not be practical in an enterprise which operates in multiple time zones.

We illustrate loading the IP network in Figure 1-4.

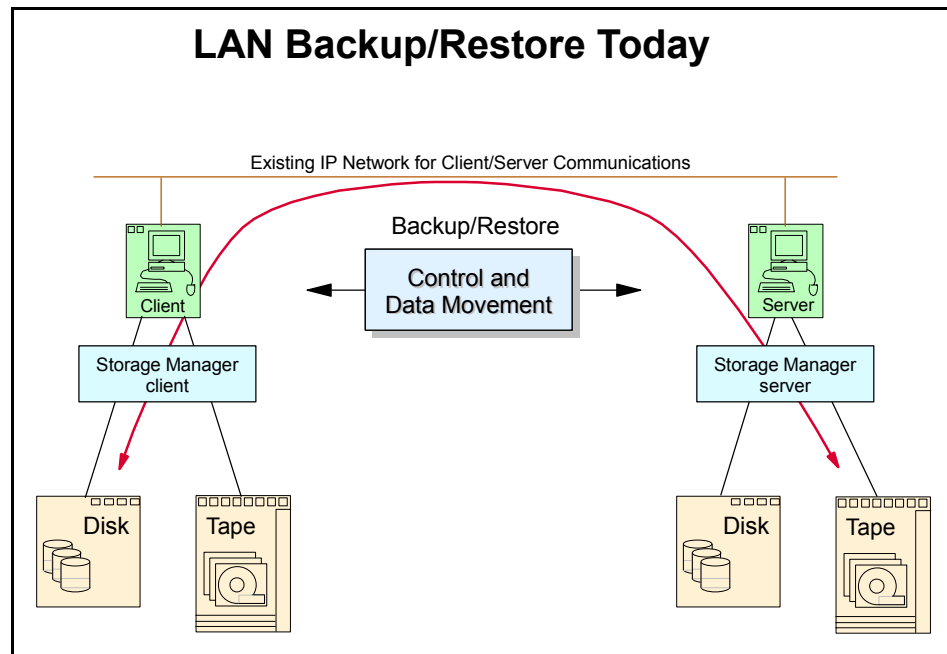


Figure 1-4 Loading the IP network

SAN provides the solution, by enabling the elimination of backup and recovery data movement across the LAN. Fibre Channel's high bandwidth and multi-path switched fabric capabilities enables multiple servers to stream backup data concurrently to high speed tape drives. This frees the LAN for other application traffic. The IBM Tivoli software solution for LAN-free backup offers the capability for clients to move data directly to tape using the SAN. A future enhancement to be provided by IBM Tivoli will allow data to be read directly from disk to tape (and tape to disk), bypassing the server. This solution is known as server-free backup.

1.3.5 High performance

Applications benefit from the more efficient transport mechanism of Fibre Channel. Currently, Fibre Channel transfers data at 200 MB/s, several times faster than typical SCSI capabilities, and many times faster than standard LAN data transfers. Future implementations of Fibre Channel at 400 and 800 MB/s have been defined, offering the promise of even greater performance benefits in the future. Indeed, prototypes of storage components which meet the 2 Gigabit transport specification are already in existence.

The elimination of conflicts on LANs, by removing storage data transfers from the LAN to the SAN, may also significantly improve application performance on servers.

1.3.6 High availability server clustering

Reliable and continuous access to information is an essential prerequisite in any business. As applications have shifted from robust mainframes to the less reliable client/file server environment, so have server and software vendors developed high availability solutions to address the exposure. These are based on clusters of servers. A cluster is a group of independent computers managed as a single system for higher availability, easier manageability, and greater scalability. Server system components are interconnected using specialized cluster interconnects, or open clustering technologies, such as Fibre Channel - Virtual Interface mapping.

Complex software is required to manage the failover of any component of the hardware, the network, or the application. SCSI cabling tends to limit clusters to no more than two servers. A Fibre Channel SAN allows clusters to scale to 4, 8, 16, and even to 100 or more servers, as required, to provide very large shared data configurations, including redundant pathing, RAID protection, and so on. Storage can be shared, and can be easily switched from one server to another. Just as storage capacity can be scaled non-disruptively in a SAN, so can the number of servers in a cluster be increased or decreased dynamically, without impacting the storage environment.

1.3.7 Improved disaster tolerance

Advanced disk arrays, such as IBM Enterprise Storage Server® (ESS), provide sophisticated functions, like Peer-to-Peer Remote Copy services, to address the need for secure and rapid recovery of data in the event of a disaster. Failures may be due to natural occurrences, such as fire, flood, or earthquake; or to human error. A SAN implementation allows multiple open servers to benefit from this type of disaster protection, and the servers may even be located some distance (up to 10 km) from the disk array which holds the primary copy of the data. The secondary site, holding the mirror image of the data, may be located up to a further 100 km from the primary site.

IBM has also announced Peer-to-Peer Copy capability for its Virtual Tape Server (VTS). This will allow VTS users to maintain local and remote copies of virtual tape volumes, improving data availability by eliminating all single points of failure.

1.3.8 Allow selection of “best of breed” storage

Internal storage, purchased as a feature of the associated server, is often relatively costly. A SAN implementation enables storage purchase decisions to be made independently of the server. Buyers are free to choose the best of breed solution to meet their performance, function, and cost needs. Large capacity external disk arrays may provide an extensive selection of advanced functions. For instance, the ESS includes cross platform functions, such as high performance RAID 5, Peer-to-Peer Remote Copy, Flash Copy, and functions specific to S/390®, such as Parallel Access Volumes (PAV), Multiple Allegiance, and I/O Priority Queuing. This makes it an ideal SAN attached solution to consolidate enterprise data.

Client/server backup solutions often include attachment of low capacity tape drives, or small automated tape subsystems, to individual PCs and departmental servers. This introduces a significant administrative overhead as users, or departmental storage administrators, often have to control the backup and recovery processes manually. A SAN allows the alternative strategy of sharing fewer, highly reliable, powerful tape solutions, such as the IBM Magstar® family of drives and automated libraries, between multiple users and departments.

1.3.9 Ease of data migration

Data can be moved non-disruptively from one storage subsystem to another using a SAN, without server intervention. This may greatly ease the migration of data associated with the introduction of new technology, and the retirement of old devices.

1.3.10 Reduced total costs of ownership

Expenditure on storage today is estimated to be in the region of 50% of a typical IT hardware budget. Some industry analysts expect this to grow to as much as 75% by the end of the year 2002. IT managers are becoming increasingly focused on controlling these growing costs.

Consistent, centralized management

As we have shown, consolidation of storage can reduce wasteful fragmentation of storage attached to multiple servers. It also enables a single, consistent data and storage resource management solution to be implemented, such as IBM StorWatch™ tools, combined with software such as Tivoli Storage Network Manager, Tivoli Storage Manager, and Tivoli SAN Manager, which can reduce costs of software and human resources for storage management.

Reduced hardware costs

By moving data to SAN attached storage subsystems, the servers themselves may no longer need to be configured with native storage. In addition, the introduction of LAN-free and server-free data transfers largely eliminate the use of server cycles to manage housekeeping tasks, such as backup and recovery, and archive and recall. The configuration of what might be termed “thin servers” therefore might be possible, and this could result in significant hardware cost savings to offset against costs of installing the SAN fabric.

1.3.11 Storage resources match e-business enterprise needs

By eliminating islands of information, typical of the client/server model of computing, and introducing an integrated storage infrastructure, SAN solutions match the strategic needs of today's e-business.

We show this in Figure 1-5.

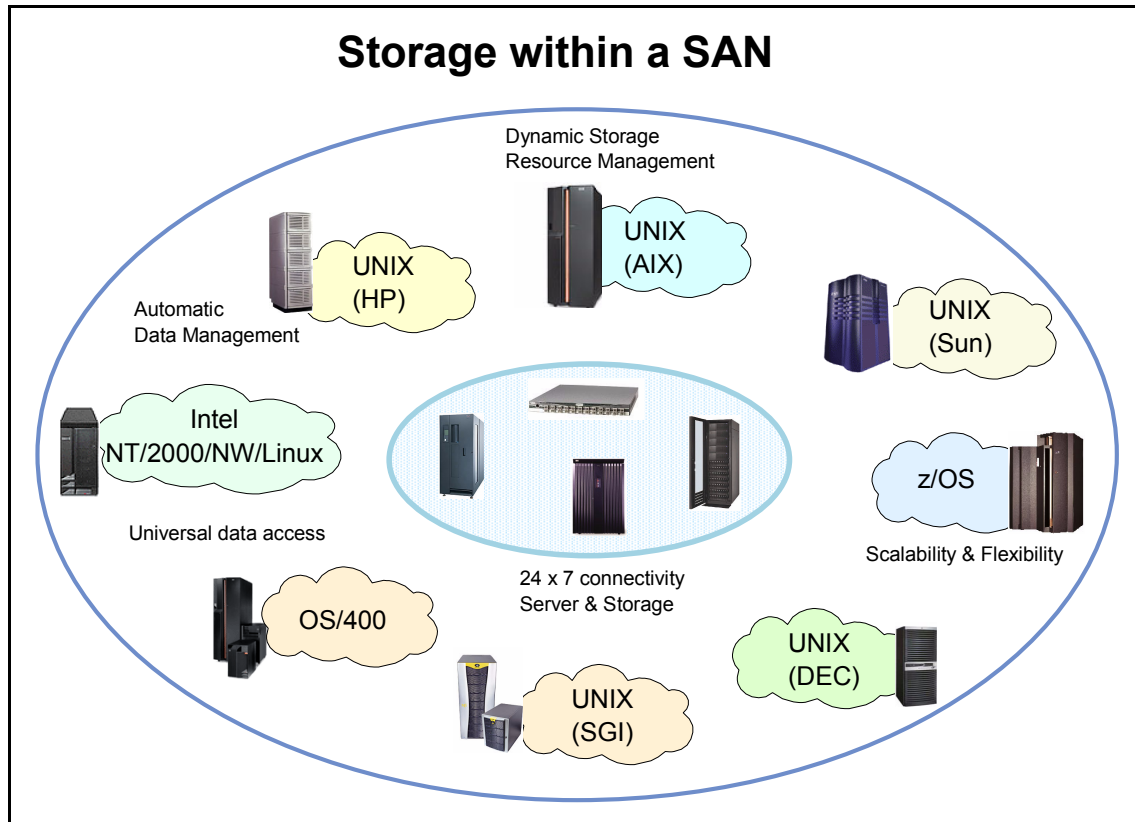


Figure 1-5 SAN total storage solutions

A well designed, well thought out SAN can bring many benefits, and not only those related to business continuance. Utilizing the storage network will be key to the storage and successful retrieval of data in the future, and the days of server centric storage are rapidly becoming a distant memory.



SAN fabric components

In this chapter we describe the Fibre Channel products that are used and are likely to be encountered in an IBM Enterprise SAN implementation. This does not mean that you cannot implement other SAN compatible products, including those from other vendors, but the interoperability agreement must be clearly documented and agreed upon.

Fibre Channel is an open standard communications and transport protocol as defined by ANSI (Committee X3T11) and operates over copper and fiber optic cabling at distances of up to 10 kilometers. IBM's implementation is in fiber optic cabling and will be referred to as Fibre Channel cabling, or FC cabling, in this redbook.

Fibre or Fiber?: Fibre Channel was originally designed to support fiber optic cabling only. When copper support was added, the committee decided to keep the name in principle but change its spelling from fiber to fibre. When referring to specific cabling, the correct American English spelling of fiber should be used.

2.1 ASIC technology

Today's hardware designers are required to balance requirements for performance and special features against steadily increasing pressure on design-cycle time, flexibility, and cost. Rapid advances in technology, library design, and design automation have made application specific integrated circuit (ASIC) technology an increasingly viable choice for many applications.

This section describes the common fabric electronics such as ASIC and its predefined set of elements, such as logic functions, I/O circuits, memory arrays and backplane.

The switch and director electronics consist of a system board and multiple port cards respectively, incorporating the Fibre Channel port interfaces, switching mechanism, embedded switch control processor, and support logic for the embedded processor logic.

The application-specific integrated circuit or commonly known as ASIC provides four Fibre Channel ports that may be used to connect to external N_Ports (such as an F_Port or FL_Port), external loop devices (such as an FL_Port), or to other switches (such as an E_Port). The ASIC contains the Fibre Channel interface logic, message/buffer queuing logic, and receive buffer memory for the on-chip ports, as well as other support logic.

2.2 Fiber optic interconnects

In Fibre Channel technology, frames are moved from source to destination using gigabit transport, which is a requirement to achieve fast transfer rates. To communicate with gigabit transport, both sides have to support this type of communication. This can be accomplished by installing this feature into the device or by using specially designed interfaces which can convert other communication transport into gigabit transport. Gigabit transport can be used in a copper or fiber optic infrastructure. We recommend that you consider using a fiber optic implementation if you need to avoid the distance limitation of SCSI, or are likely to in the future.

Nowadays, fibre optic implementations are much more common due to the ease of use and flexibility that it offers. With earlier Fibre Channel implementations, the cost of the copper infrastructure was more attractive, but this is not as significant an issue today.

The interfaces that are used to convert the internal communication transport of gigabit transport are as follows:

- ▶ Small Form Factor Transceivers (SFF)
- ▶ Gigabit Interface Converters (GBIC)
- ▶ Gigabit Link Modules (GLM)
- ▶ 1x9 transceivers
- ▶ Media Interface Adapters (MIA)

We provide a brief description of the types of cables and connectors, and their functions in the following topics.

2.2.1 Small Form Factor Optical Transceivers

The IBM 1063 Mb/s up to 2125 Mb/s Small Form Factor (SFF) serial optical converters are the next generation of laser-based, optical transceivers for a wide range of networking applications requiring high data rates. The transceivers, which are designed for increased densities, performance, and reduced power, are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications (see Figure 2-1).

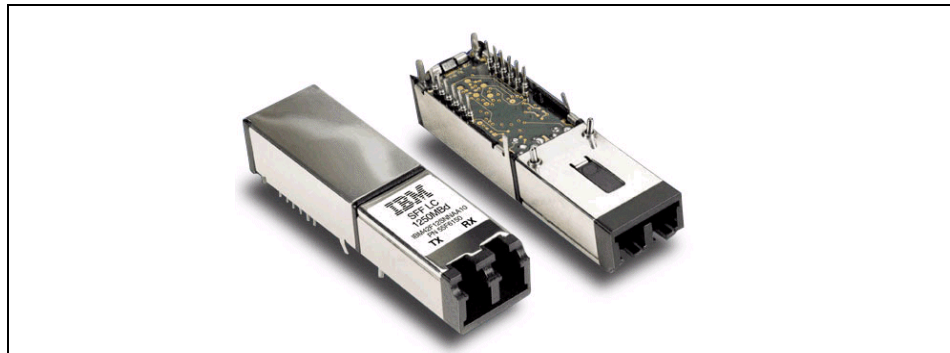


Figure 2-1 Small Form Factor Transceiver

The SFF optical transceivers use short wavelength and long wavelength lasers and are available in pin-through-hole (PTH) or hot-pluggable versions.

The small dimensions of the SFF optical transceivers are ideal in switches and other products where many transceivers have to be configured in a small space. Using these SFF devices, manufacturers can increase the density of transceivers on a board, compared with what was possible using previous optical transceiver technologies. This device is flexible, self-configuring for 100 MB/s or 200 MB/s transmission rates for current or future speeds, providing a seamless transition. Its enhanced design features include frequency agility, reduced power consumption, and lower cost transmission.

The SFF serial optical transceivers are integrated fiber-optic transceivers providing a high-speed serial electrical interface for connecting processors, switches, and peripherals through an optical fiber cable. In the Gigabit Ethernet environment, for example, these transceivers can be used in local area network (LAN) switches or hubs, as well as in interconnecting processors. In SANs, they can be used for transmitting data between peripheral devices and processors.



The distances that can be achieved using the SFF short wavelength and long wavelength are listed in Table 2-1.

Type of fiber	SWL	LWL
9/125 μm Optical Fiber	n/a	<ul style="list-style-type: none"> Distance for 100 MB/s version: 2 m - 10 km Distance for 1.250 Gb/s version: 2 m - 5 km Distance for 2.125 Gb/s version: 2 m - 10 km

Type of fiber	SWL	LWL
50/125 μ m Optical Fiber	<ul style="list-style-type: none"> Distance for 1.0625 Gb/s version: 2 - 500 m Distance for 1.250 Gb/s version: 2 - 550 m Distance for 2.125 Gb/s version: 2 - 300 m 	2 - 550 m
62.5/125 μ m Optical Fiber	<ul style="list-style-type: none"> Distance for 1.0625 Gb/s version: 2 - 300 m Distance for 1.250 Gb/s version: 2 - 275 m Distance for 2.125 Gb/s version: 2 - 150 m 	2 - 550 m

The distances shown are not necessarily the supported distances, and this will have to be verified with the fiber transport installer.

SFP: The Small Form Factor Hot-Pluggable module is also known as SFP. The IBM SFF optical transceiver is offered with an LC connector due to its robustness, increased transmission performance, better EMI, virtually no optical or electrical cross talk, and less jitter, with the primary objective being to ensure reliability, serviceability, and availability.

An SFF is not always hot-pluggable, whereas an SFP is.

2.2.2 Gigabit Interface Converters

The IBM 1063 Mb/s and 1250 Mb/s Gigabit Interface Converters (GBICs) are laser-based, hot-pluggable, data communications transceivers for a wide range of networking applications requiring high data rates. The transceivers, which are designed for ease of configuration and replacement, are well-suited for Gigabit Ethernet, Fibre Channel, and 1394b applications.

The 1063 Mb/s GBIC and 1250 Mb/s GBIC are available in both short wavelength and long wavelength versions, providing configuration flexibility. Users can easily add a GBIC in the field to accommodate a new configuration requirement or replace an existing device to allow for increased availability.

The 1063 Mb/s GBIC and 1250 Mb/s GBIC are integrated fiber-optic transceivers providing a high-speed serial electrical interface for connecting processors, switches, and peripherals through an optical fiber cable. In the Gigabit Ethernet environment, for example, these transceivers can be used in local area network (LAN) switches or hubs, as well as in interconnecting processors. In SANs they can be used for transmitting data between peripheral devices and processors.

The GBICs use lasers that enable cost-effective data transmission over optical fibers at distances of up to 10 km. These compact, hot-pluggable, field-replaceable modules are designed to connect easily to a system card through an industry-standard connector. Single-mode or multi-mode optical fiber cables, terminated with industry-standard SC connectors, can be used.

Also available now are 2125 Mb/s and 2500 Mb/s GBICs (referred to below as 2 Gb/s GBIC). We have included below the distance capabilities for each of the different types of GBICs.

There are two types of connections on the service side, namely, DB-9 and HSSDC. The fiber-optic type has two modes of operation. The difference between them is in the laser wave length. The two modes are:

- ▶ Short wavelength mode: SWL
- ▶ Long wavelength mode: LWL

The distances that can be achieved using both the 1 Gb/s-based GBICs are shown in Table 2-2.

Table 2-2 Distance using 1 Gb/s GBIC based fiber optics

Type of fiber	SWL	LWL
9/125 μ m Optical Fiber	n/a	10 km
50/125 μ m Optical Fiber	2 - 550 m	2 - 550 m
62.5/125 μ m Optical Fiber	2 - 300 m	2 - 550 m

The distances that can be achieved using both the 2Gb/s based GBICs are shown in Table 2-3.

Table 2-3 Distance using 2 Gb/s GBIC based fiber-optics

Type of fiber	SWL	LWL
9/125 μ m Optical Fiber	n/a	10 km
50/125 μ m Optical Fiber	2 - 300 m	2 - 550 m
62.5/125 μ m Optical Fiber	2 - 150 m	2 - 550 m

The standard dual SC plug is used to connect to the fiber optic cable. This is shown in Figure 2-3.

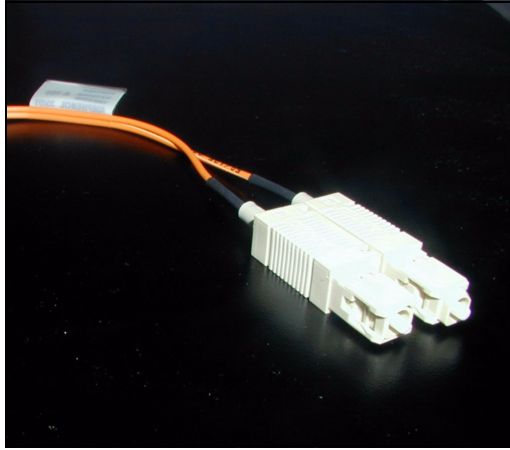


Figure 2-3 Dual SC fiber-optic plug connector

GBICs are usually hot-pluggable, easy to configure and replace. On the optical side they use low-loss, SC type, push-pull, optical connectors. They are mainly used in hubs, switches, directors, and gateways.

Shortwave (or multi-mode) GBICs are usually color coded beige with a black exposed surface; and longwave (or single-mode) GBICs are usually color coded blue with blue exposed surfaces.

The transfer rates typically range from 1063 Mb/s, 1250 Mb/s, 2125 Mb/s, or 2500 Mb/s. A GBIC is shown in Figure 2-4.



Figure 2-4 Gigabit Interface Converter

The selection of a GBIC for SAN interconnection is just as important a consideration as choosing a hub or a switch, and should not be overlooked or taken lightly.

2.2.3 Gigabit Link Modules

Gigabit Link Modules (GLMs) — sometimes referred to as Gigabaud Link Modules) — were used in early Fibre Channel applications. GLMs are a low cost alternative to GBICs, but they sacrifice the ease of use and hot-pluggable installation and replacement characteristics that GBICs offer. This means that you need to power down the device for maintenance, replacement, or repair.

GLMs enable computer manufacturers to integrate low-cost, high-speed fiber optic communications into devices. They use the same fiber optic for the transport of optical signal as GBICs. GLMs also use two types of lasers, SWL and LWL, to transport the information across the fiber optic channel. The transfer rates that are available are 266 Mb/s and 1063 Mb/s.

The 266 Mb/s and 1063 Mb/s GLM cards support continuous, full-duplex communication. The GLM converts encoded data that has been serialized into pulses of laser light for transmission into the optical fiber. A GLM at a second optical link, running at the same speed as the sending GLM, receives these pulses, along with the requisite synchronous clocking signals.

With 1063 Mb/s you can achieve the distances listed in Table 2-2, “Distance using 1 Gb/s GBIC based fiber optics” on page 24.

A GLM is shown in Figure 2-5.

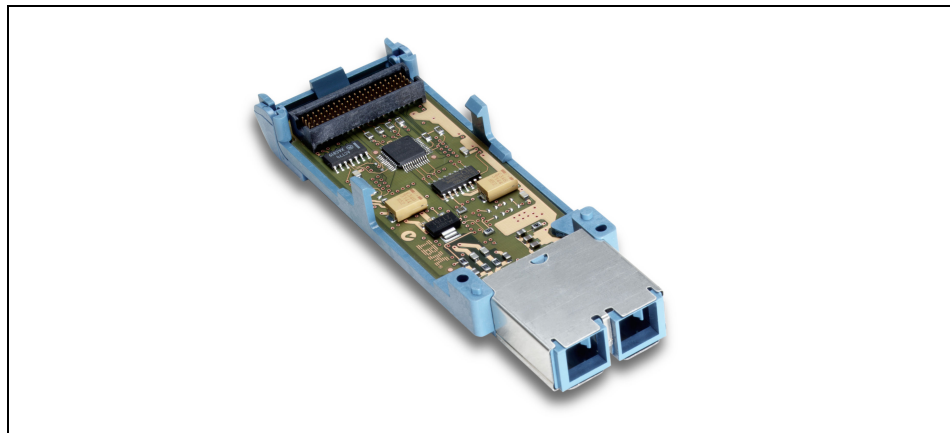


Figure 2-5 Gigabit Link Module

2.2.4 Media Interface Adapters

Media Interface Adapters (MIAs) can be used to facilitate conversion between optical and copper interface connections. Typically, MIAs are attached to host bus adapters, but they can also be used with switches and hubs. If a hub or switch only supports copper or optical connections, MIAs can be used to convert the signal to the appropriate media type, copper or optical.

An MIA is shown in Figure 2-6.



Figure 2-6 Media Interface Adapter

2.2.5 1x9 transceivers

Some switch manufacturers prefer to use 1x9 transceivers for providing SC connection to their devices. 1x9 transceivers (Figure 2-7) have some advantages over GBICs:

- ▶ Easier to cool
- ▶ Better air flow
- ▶ More reliable (2.5 times that of a GBIC)

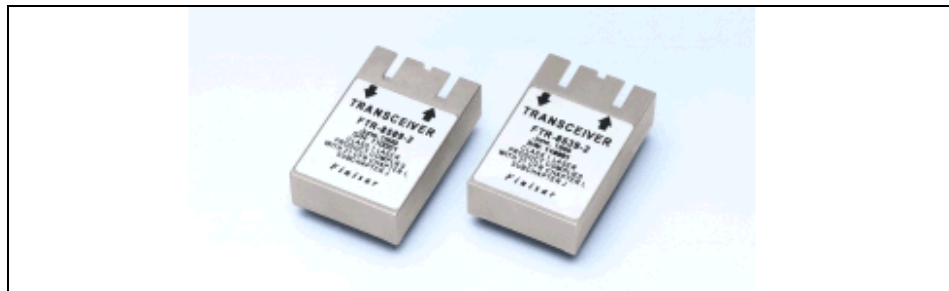


Figure 2-7 1x9 transceivers

2.2.6 Fibre Channel adapter cable

The LC-SC adapter cable attaches to the end of an LC-LC cable to support SC device connections. A combination of one LC/LC fiber cable and one LC/SC adapter cable is required for each connection. This is used to connect from some of the older 1 Gb/s devices to a 2 Gb/s capable and LC interface-based SAN.

Shown in Figure 2-6 is a Fibre Channel adapter cable.

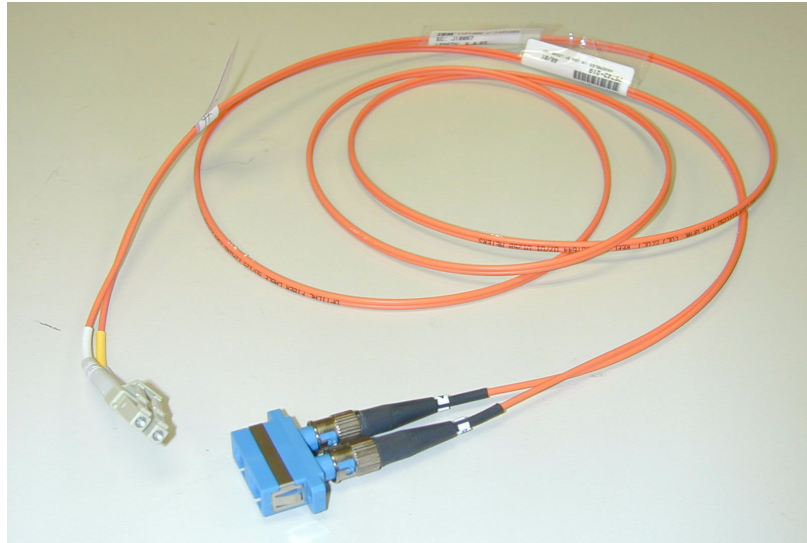


Figure 2-8 Fibre Channel adapter cable

2.3 Fibre Channel ports

Before we look at the concepts that make up the SAN topology, we first introduce the terminology that is important to understand when joining a discussion on SAN. At various stages throughout this chapter, we then begin to describe some of the most common items that are met.

2.3.1 Port types

These are the types of Fibre Channel ports that are likely to be encountered:

- ▶ **E_Port:** This is an expansion port. A port is designated an E_Port when it is used as an inter switch expansion port (ISL) to connect to the E_Port of another switch, to enlarge the switch fabric.
- ▶ **F_Port:** This is a fabric port that is not loop capable. It is used to connect an N_Port point-to-point to a switch.
- ▶ **FL_Port:** This is a fabric port that is loop capable. It is used to connect NL_Ports to the switch in a public loop configuration.
- ▶ **G_Port:** This is a generic port that can operate as either an E_Port or an F_Port. A port is defined as a G_Port after it is connected but has not received response to *loop* initialization or has not yet completed the *link* initialization procedure with the adjacent Fiber Channel device.
- ▶ **L_Port:** This is a loop capable node or switch port.
- ▶ **U_Port:** This is a universal port. A more generic switch port than a G_Port. It can operate as either an E_Port, F_Port, or FL_Port. A port is defined as a U_Port when it is not connected or has not yet assumed a specific function in the fabric.
- ▶ **N_Port:** This is a node port that is not loop capable. It is used to connect an equipment port to the fabric.
- ▶ **NL_Port:** This is a node port that is loop capable. It is used to connect an equipment port to the fabric in a loop configuration through an L_Port or FL_Port.
- ▶ **TL_Port:** This is a private to public bridging of switches or directors (the Cisco Systems MDS 9000 Family also uses translatable loop ports).

Figure 2-9 represents the most commonly encountered Fibre Channel port types.

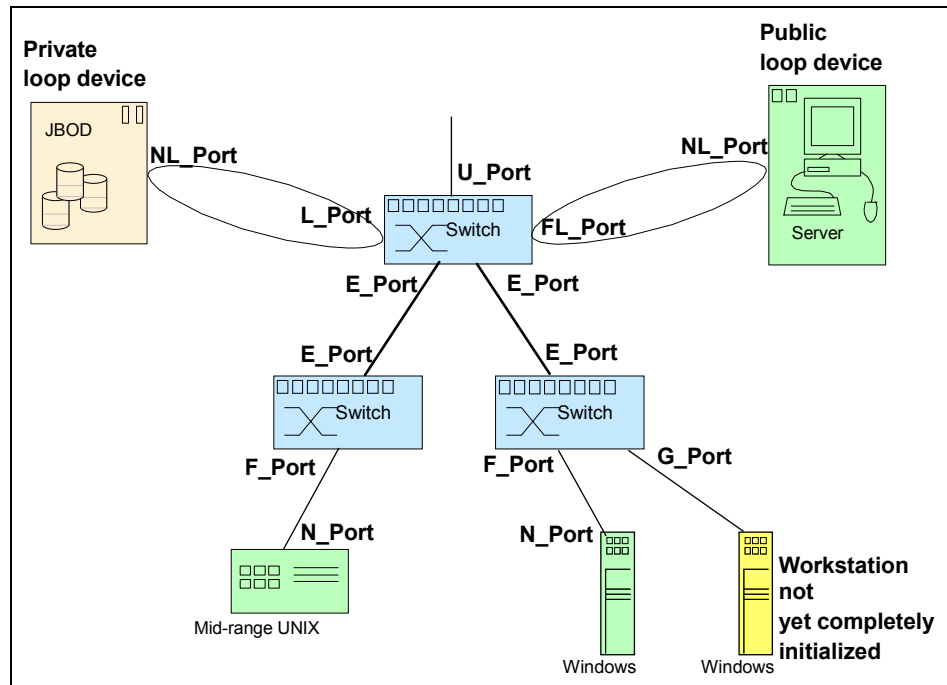


Figure 2-9 Fibre Channel port types

2.4 SAN topologies

Fibre Channel provides three distinct interconnection topologies. By having more than one interconnection option available, a particular application can choose the topology that is best suited to its requirements. The three Fibre Channel topologies are:

- ▶ Point-to-point
- ▶ Arbitrated loop
- ▶ Switched fabric

We discuss these in greater detail in the topics that follow.

2.4.1 Point-to-point

A point-to-point connection is the simplest topology. It is used when there are exactly two nodes, and future expansion is not predicted. There is no sharing of the media, which allows the devices to use the total bandwidth of the link. A simple link initialization is needed before communications can begin.

Fibre Channel is a full duplex protocol, which means both paths simultaneously. Fibre Channel connections based on the 1 Gb standard are able to transmit at 100 MB/s and receive at 100 MB/s simultaneously.

For Fibre Channel connections based on the 2 Gb standard, they can transmit at 200 MB/s and receive at 200 MB/s simultaneously.

Illustrated in Figure 2-10 is a simple point-to-point connection.

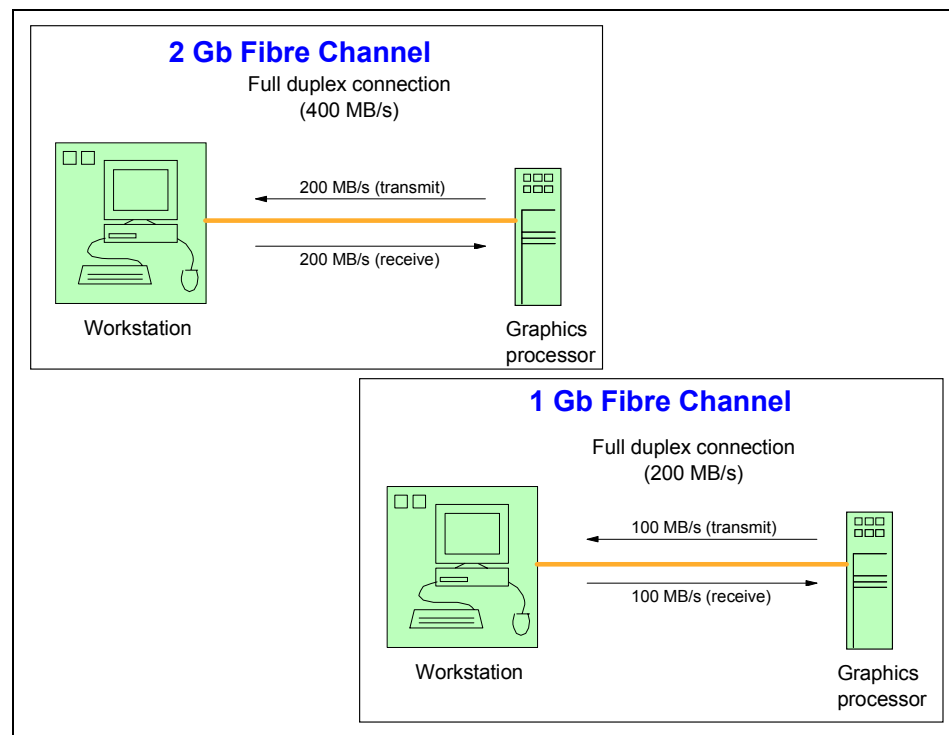


Figure 2-10 Point-to-point

An extension of the point-to-point topology is the logical start topology. This is a collection of point-to-point topology links and both topologies provide full duplex bandwidth.

2.4.2 Arbitrated loop

The second topology is Fibre Channel Arbitrated Loop (FC-AL). FC-AL is more useful for storage applications. It is a loop of up to 126 nodes (NL_Ports) that is managed as a shared bus. Traffic flows in one direction, carrying data frames and primitives around the loop with a total bandwidth of 200 MB/s (or 100 MB/s for a loop based on 1 Gb/s technology).

Using arbitration protocol, a single connection is established between a sender and a receiver, and a data frame is transferred around the loop. When the communication comes to an end between the two connected ports, the loop becomes available for arbitration and a new connection may be established. Loops can be configured with hubs to make connection management easier. A distance of up to 10 km is supported by the Fibre Channel standard for both of these configurations. However, latency on the arbitrated loop configuration is affected by the loop size.

A simple loop, configured using a hub, is shown in Figure 2-11.

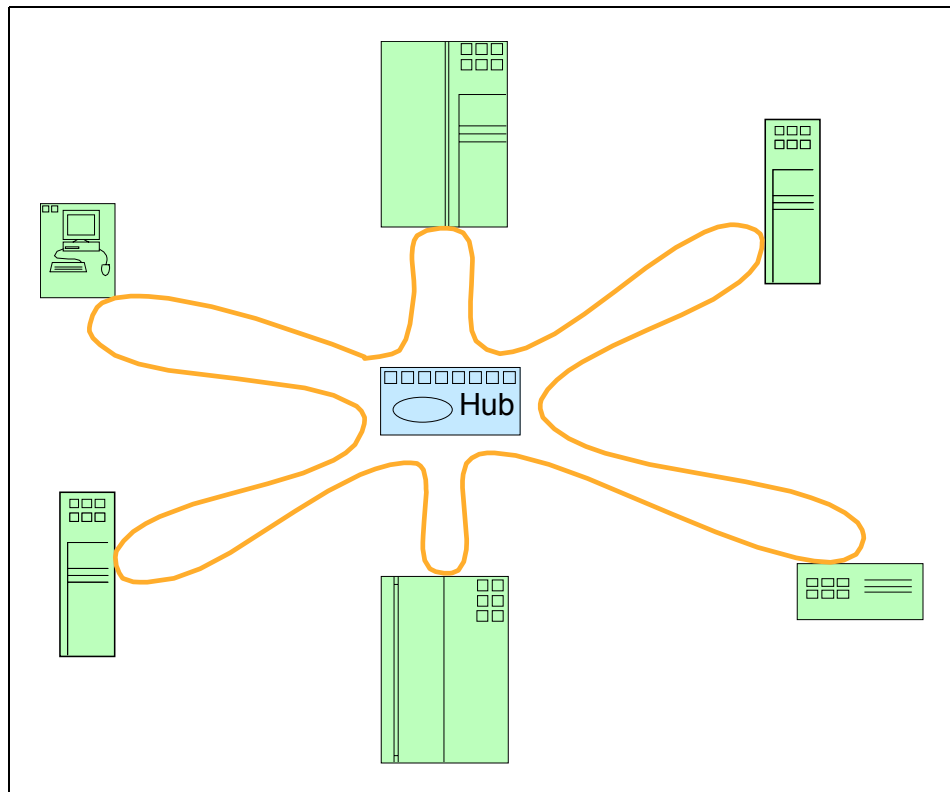


Figure 2-11 Arbitrated loop

Loop protocols

To support the shared behavior of the arbitrated loop, a number of loop-specific protocols are used. These protocols are used to:

- ▶ Initialize the loop and assign addresses.
- ▶ Arbitrate for access to the loop.
- ▶ Open a loop circuit with another port in the loop.
- ▶ Close a loop circuit when two ports have completed their current use of the loop.
- ▶ Implement the access fairness mechanism to ensure that each port has an opportunity to access the loop.

Loop initialization

Loop initialization is a necessary process for the introduction of new participants on to the loop. Whenever a loop port is powered on or initialized, it executes the loop initialization primitive (LIP) to perform loop initialization. Optionally, loop initialization may build a positional map of all the ports on the loop. The positional map provides a count of the number of ports on the loop, their addresses and their position relative to the loop initialization master.

Following loop initialization, the loop enters a stable monitoring mode and begins with normal activity. An entire loop initialization sequence may take only a few milliseconds, depending on the number of NL_Ports attached to the loop. Loop initialization may be started by a number of causes. One of the most likely reasons for loop initialization is the introduction of a new device. For instance, an active device may be moved from one hub port to another hub port, or a device that has been powered on could re-enter the loop.

A variety of ordered sets have been defined to take into account the conditions that an NL_Port may sense as it starts the initialization process. These ordered sets are sent continuously while a particular condition or state exists. As part of the initialization process, loop initialization primitive sequences (referred to collectively as LIPs) are issued. As an example, an NL_Port must issue at least three identical ordered sets to start initialization. An ordered set transmission word always begins with the special character K28.5.

Once these identical ordered sets have been sent, and as each downstream device receives the LIP stream, devices enter a state known as open-init. This causes the suspension of any current operation and enables the device for the loop initialization procedure. LIPs are forwarded around the loop until all NL_Ports are in an open-init condition.

At this point, the NL_Ports need to be managed. In contrast to a Token-Ring, the Arbitrated Loop has no permanent master to manage the topology.

Therefore, loop initialization provides a selection process to determine which device will be the temporary loop master. After the loop master is chosen it assumes the responsibility for directing or managing the rest of the initialization procedure. The loop master also has the responsibility for closing the loop and returning it to normal operation.

Selecting the loop master is carried out by a subroutine known as the Loop Initialization Select Master (LISM) procedure. A loop device can be considered for temporary master by continuously issuing LISM frames that contain a port type identifier and a 64-bit World-Wide Name. For FL_Ports the identifier is x'00' and for NL_Ports it is x'EF'.

When a downstream port receives a LISM frame from an upstream partner, the device will check the port type identifier. If the identifier indicates an NL_Port, the downstream device will compare the WWN in the LISM frame to its own. The WWN with the lowest numeric value has priority. If the received frame's WWN indicates a higher priority, that is to say it has a lower numeric value, the device stops its LISM broadcast and starts transmitting the received LISM. Had the received frame been of a lower priority, the receiver would have thrown it away and continued broadcasting its own.

At some stage in proceedings, a node will receive its own LISM frame, which indicates that it has the highest priority, and succession to the throne of temporary loop master has taken place. This node will then issue a special ordered set to indicate to the others that a temporary master has been selected.

Hub cascading

Since an arbitrated loop hub supplies a limited number of ports, building larger loops may require linking another hub. This is called hub cascading. A server with an FC-AL, shortwave, host bus adapter can connect to an FC-AL hub 500 m away. Each port on the hub can connect to an FC-AL device up to 500 m away. Cascaded hubs use one port on each hub for the hub-to-hub connection and this increases the potential distance between nodes in the loop by an additional 500 m. In this topology the overall distance is 1500 m. Both hubs can support other FC-AL devices at their physical locations. Stated distances assume a 50 micron multi-mode cable.

Loops

There are two different kinds of loops: private and public.

- **Private loop:** The private loop does not connect with a fabric, only to other private nodes using attachment points called NL_Ports. A private loop is enclosed and known only to itself. In Figure 2-12 we show a private loop.

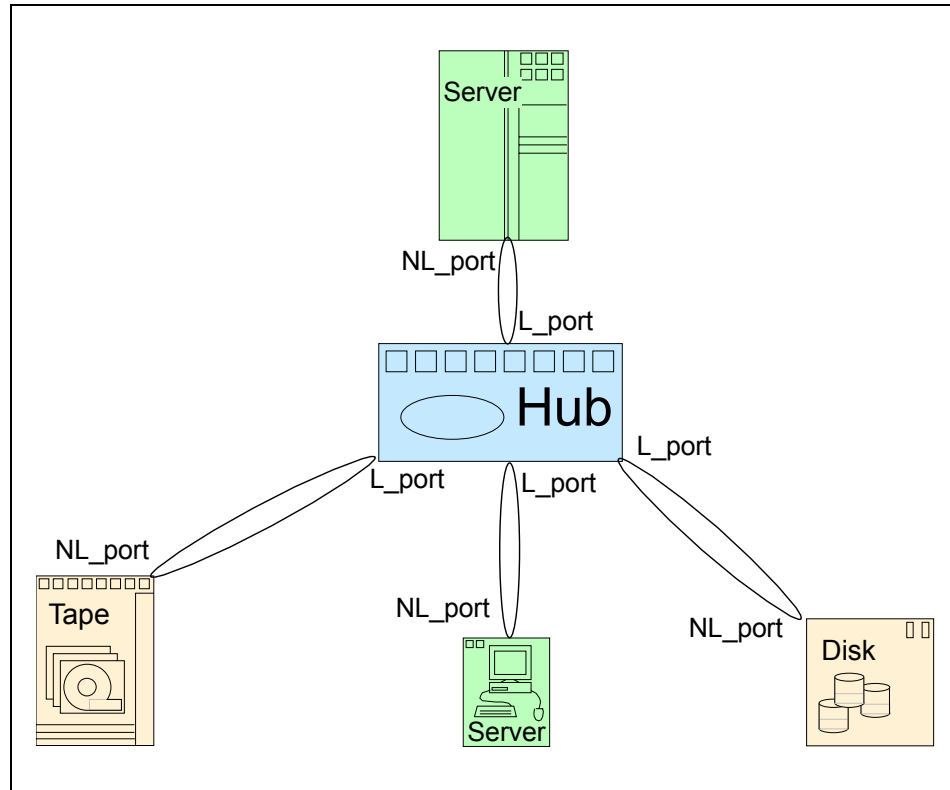


Figure 2-12 Private loop implementation

- **Public loop:** A public loop requires a fabric and has at least one FL_Port connection to a fabric. A public loop extends the reach of the loop topology by attaching the loop to a fabric. Figure 2-13 shows a public loop.

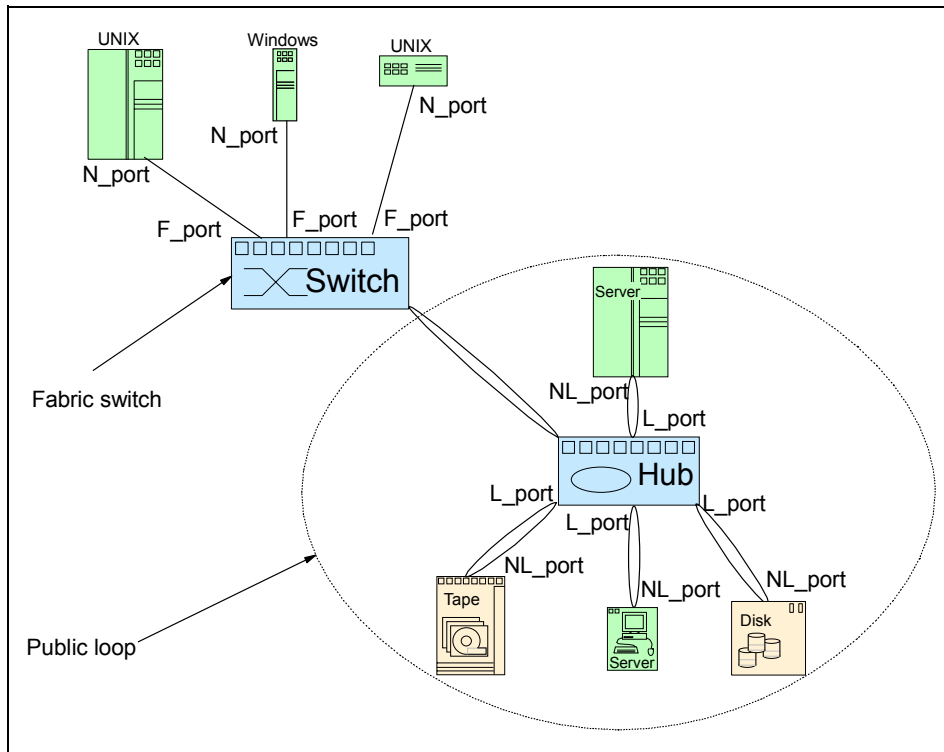


Figure 2-13 Public loop implementation

Arbitration

When a loop port wants to gain access to the loop, it has to arbitrate. When the port wins arbitration, it can open a loop circuit with another port on the loop; a function similar to selecting a device on a bus interface. Once the loop circuit has been opened, the two ports can send and receive frames between each other. This is known as loop tenancy.

If more than one node on the loop is arbitrating at the same time, the node with the lower Arbitrated Loop Physical Address (AL_PA) gains control of the loop. Upon gaining control of the loop, the node then establishes a point-to-point transmission with another node using the full bandwidth of the media. When a node has finished transmitting its data, it is not required to give up control of the loop. This is a channel characteristic of Fibre Channel. However, there is a fairness algorithm, which states that a device cannot regain control of the loop until the other nodes have had a chance to control the loop.

Loop addressing

An NL_Port, like an N_Port, has a 24-bit port address. If no switch connection exists, the two upper bytes of this port address are zeroes (x'00 00') and referred to as a private loop. The devices on the loop have no connection with the outside world. If the loop is attached to a fabric and an NL_Port supports a fabric login, the upper two bytes are assigned a positive value by the switch. We call this mode a public loop.

As fabric-capable NL_Ports are members of both a local loop and a greater fabric community, a 24-bit address is needed as an identifier in the network. In the case of public loop assignment, the value of the upper two bytes represents the loop identifier, and this will be common to all NL_Ports on the same loop that performed login to the fabric.

In both public and private arbitrated loops, the last byte of the 24-bit port address refers to the arbitrated loop physical address (AL_PA). The AL_PA is acquired during initialization of the loop and may, in the case of fabric-capable loop devices, be modified by the switch during login.

The total number of the AL_PAs available for arbitrated loop addressing is 127. This number is based on the requirements of 8b/10b running disparity between frames.

As a frame terminates with an end-of-frame character (EOF), this will force the current running disparity negative. In the Fibre Channel standard each transmission word between the end of one frame and the beginning of another frame should also leave the running disparity negative. If all 256 possible 8-bit bytes are sent to the 8b/10b encoder, 134 emerge with neutral disparity characters. Of these 134, seven are reserved for use by Fibre Channel. The 127 neutral disparity characters left have been assigned as AL_PAs. Put another way, the 127 AL_PA limit is simply the maximum number, minus reserved values, of neutral disparity addresses that can be assigned for use by the loop. This does not imply that we recommend this amount, or load, for a 200MB/s shared transport, but only that it is possible.

Arbitrated loop will assign priority to AL_PAs, based on numeric value. The lower the numeric value, the higher the priority is. For example, an AL_PA of x'01' has a much better position to gain arbitration over devices that have a lower priority or higher numeric value. At the top of the hierarchy it is not unusual to find servers, but at the lower end you would expect to find disk arrays.

It is the arbitrated loop initialization that ensures each attached device is assigned a unique AL_PA. The possibility for address conflicts only arises when two separated loops are joined together without initialization.

2.4.3 Logins

There are three different types of logins for Fibre Channel:

- ▶ Fabric login
- ▶ Port login
- ▶ Process login

Here we describe only the port login and the process login. Later, we provide details on the fabric login; you can refer to “Fabric login” on page 44.

Port login

The port login is also known as PLOGI.

A port login is used to establish a session between two N_Ports (devices) and is necessary before any upper level commands or operations can be performed. During the port login, two N_Ports (devices) swap service parameters and make themselves known to each other.

Process login

The process login is also known as PRLI. The process login is used to set up the environment between related processes on an originating N_Port and a responding N_Port. A group of related processes is collectively known as an image pair. The processes involved can be system processes, system images, such as mainframe logical partitions, control unit images, and FC-4 processes. Use of the process login is optional from the perspective of the Fibre Channel FC-2 layer, but may be required by a specific upper-level protocol, as in the case of SCSI-FCP mapping.

We show the Fibre Channel logins in Figure 2-14.

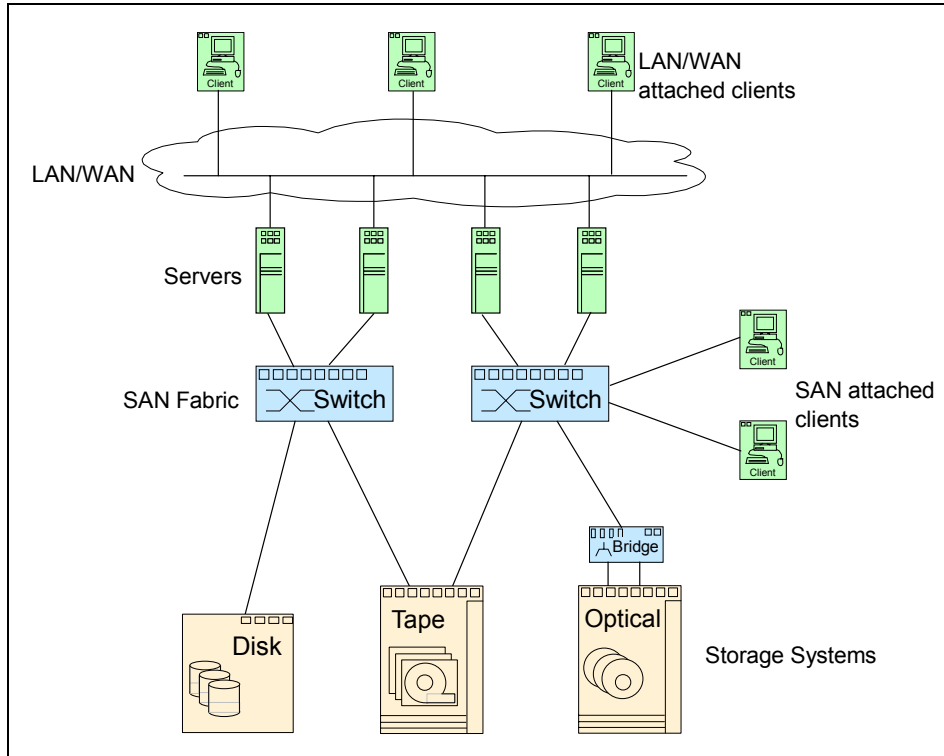


Figure 2-14 Fibre Channel logins

Closing a loop circuit

When two ports in a loop circuit complete their frame transmission, they may close the loop circuit to allow other ports to use the loop. The point at which the loop circuit is closed depends on the higher-level protocol, the operation in progress, and the design of the loop ports.

Supported devices

An arbitrated loop may support a variety of devices, such as:

- ▶ Individual Fibre Channel disk drives
- ▶ JBOD
- ▶ Fibre Channel RAID
- ▶ Native Fibre Channel tape sub-systems
- ▶ Fibre Channel to SCSI bridges

Broadcast

Arbitrated loop, in contrast to Ethernet, is a non-broadcast transport. When an NL_Port has successfully won the right to arbitration, it will open a target for frame transmission. Any subsequent loop devices in the path between the two will see the frames and forward them on to the next node in the loop.

It is this non-broadcast nature of arbitrated loop, by removing frame handling overhead from some of the loop, which enhances performance.

Distance

As stated before, arbitrated loop is a closed-ring topology. The total distance requirements being determined by the distance between the nodes. At gigabit speeds, signals propagate through fiber-optic media at five nanoseconds per meter and through copper media at four nanoseconds per meter. This is the delay factor.

Calculating the total propagation delay incurred by the loop's circumference is achieved by multiplying the length — both transmit and receive — of copper and fiber-optic cabling deployed by the appropriate delay factor. For example, a single 10 km link to an NL_Port would cause a 50 microsecond (10 km x 5 nanoseconds delay factor) propagation delay in each direction and 100 microseconds in total. This equates to 1 MB/s of bandwidth used to satisfy the link.

Bandwidth

For optical interconnects for SANs, the bandwidth requirements are greatly influenced by the capabilities of:

- ▶ The system buses
- ▶ Network switches
- ▶ The interface adapters that interface with them
- ▶ Traffic locality

The exact bandwidth required is somewhat dependent on implementation, but is currently in the range of 100 to 1000 MB/s. Determining bandwidth requirements is difficult, and there is no exact science that can take into account the unpredictability of sporadic bursts of data, for example. Planning bandwidth based on peak requirements could be wasteful. Designing for sustained bandwidth requirements, with the addition of safety margins, may be less wasteful.

2.4.4 Switched fabric

The third topology used in SAN implementations is Fibre Channel Switched Fabric (FC-SW). It applies to directors that support the FC-SW standard, that is, it is not limited to switches as its name suggests. A Fibre Channel fabric is one or more fabric switches in a single, sometimes extended, configuration. Switched fabrics provide full 200 MB/s bandwidth per port (or 100 MB/s for devices based on the older 1 Gb/s infrastructure), compared to the shared bandwidth per port in Arbitrated loop implementations.

If you add a new device into the arbitrated loop, you further divide the shared bandwidth. However, in a switched fabric, adding a new device or a new connection between existing ones actually increases the bandwidth. For example, an 8-port switch (based on 2 Gb/s technology) with three initiators and three targets can support three concurrent 200 MB/s conversations or a total of 600 MB/s throughput (1,200 MB/s if full-duplex applications were available).

A switched fabric configuration is shown in Figure 2-15.

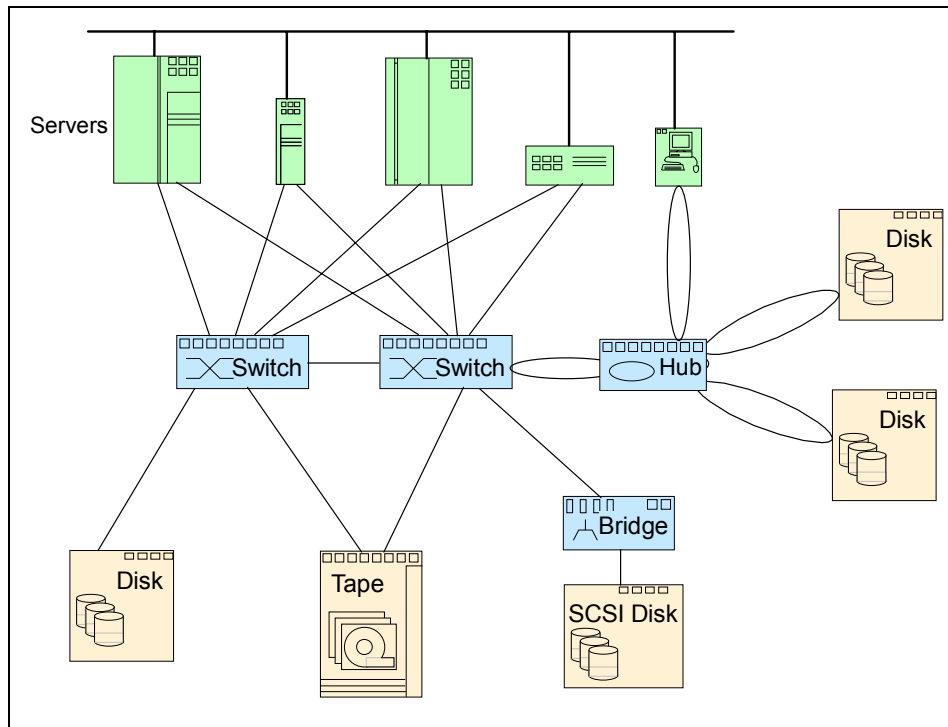


Figure 2-15 Sample switched fabric configuration

Addressing

As we know from “Name and addressing” on page 42, each participant in the Fibre Channel environment has a unique ID, which is called the World Wide Name (WWN). This WWN is a 64-bit address, and if two WWN addresses are put into the frame header, this leaves 16 bytes of data just for identifying destination and source address. So 64-bit addresses can impact routing performance.

Because of this, there is another addressing scheme used in Fibre Channel networks. This scheme is used to address the ports in the switched fabric. Each port in the switched fabric has its own unique 24-bit address. With this 24-bit addressing scheme, we get a smaller frame header, and this can speed up the routing process. With this frame header and routing logic, the Fibre Channel fabric is optimized for high-speed switching of frames.

With a 24-bit addressing scheme, this allows for up to 16 million addresses, which is an address space larger than any practical SAN design in existence in today's world. This 24-bit address has to somehow be connected to and with the 64-bit address associated with World Wide Names. We explain how this works in the following section.

Name and addressing

The 24-bit address scheme also removes the overhead of manual administration of addresses by allowing the topology itself to assign addresses. This is not like WWN addressing, in which the addresses are assigned to the manufacturers by the IEEE standards committee, and are built in to the device at build time, similar to naming a child at birth. If the topology itself assigns the 24-bit addresses, then somebody has to be responsible for the addressing scheme from WWN addressing to port addressing.

In the switched fabric environment, the switch itself is responsible for assigning and maintaining the port addresses. When the device with its WWN is logging into the switch on a specific port, the switch will assign the port address to that port, and the switch will also maintain the correlation between the port address and the WWN address of the device on that port. This function of the switch is implemented by using a name server.

The name server is a component of the fabric operating system, which runs inside the switch. It is essentially a database of objects in which fabric-attached device registers its values.

Dynamic addressing also removes the potential element of human error in address maintenance, and provides more flexibility in additions, moves, and changes in the SAN.

Port address

A 24-bit port address consists of three parts:

- ▶ Domain (bits from 23 to 16)
- ▶ Area (bits from 15 to 08)
- ▶ Port or arbitrated loop physical address — AL_PA (bits from 07 to 00)

We show how the address is built up in Figure 2-16.

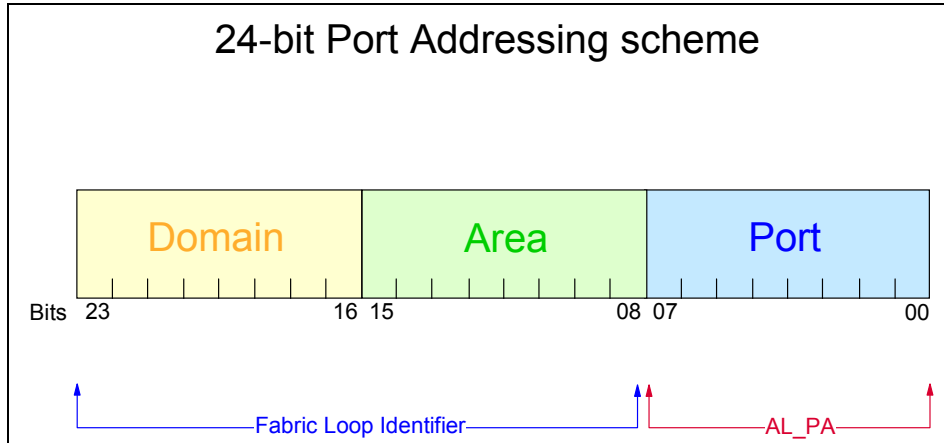


Figure 2-16 Fabric port address

Next we explain the significance of some of the bits that make up the port address:

- ▶ **Domain:** The most significant byte of the port address is the domain. This is the address of the switch itself. One byte allows up to 256 possible addresses. Because some of these are reserved (like the one for broadcast) there are only 239 addresses actually available. This means that you can have as many as 239 switches in your SAN environment. The domain number allows each switch to have a unique identifier if you have multiple interconnected switches in your environment.
- ▶ **Area:** The area field provides 256 addresses. This part of the address is used to identify the individual FL_Ports supporting loops, or it can be used as the identifier for a group of F_Ports; for example, a card with more ports on it. This means that each group of ports has a different area number, even if there is only one port in the group.
- ▶ **Port:** The final part of the address provides 256 addresses for identifying attached N_Ports and NL_Ports.

To arrive at the number of available addresses is a simple calculation based on:

Domain x Area x Ports

This means that there are $239 \times 256 \times 256 = 15,663,104$ addresses available.

Fabric login

After the fabric capable Fibre Channel device is attached to a fabric switch, it will carry out a fabric login (FLOGI).

Similar to port login, FLOGI is an extended link service command that sets up a session between two participants. With FLOGI, a session is created between an N_Port or NL_Port and the switch. An N_Port will send a FLOGI frame that contains its Node Name, its N_Port Name, and service parameters to a well-known address of 0xFFFFFE.

A public loop NL_Port first opens the destination AL_PA 0x00 before issuing the FLOGI request. In both cases the switch accepts the login and returns an accept (ACC) frame to the sender. If some of the service parameters requested by the N_Port or NL_Port are not supported, the switch will set the appropriate bits in the ACC frame to indicate this.

When the N_Port logs in, it uses a 24-bit port address of 0x000000. Because of this, the fabric is allowed to assign the appropriate port address to that device, based on the Domain-Area-Port address format. The newly assigned address is contained in the ACC response frame.

When the NL_Port logs in, a similar process starts, except that the least significant byte is used to assign AL_PA, and the upper two bytes constitute a fabric loop identifier. Before an NL_Port logs in, it will go through the LIP on the loop, which is started by the FL_Port, and from this process it has already derived an AL_PA. The switch then decides if it will accept this AL_PA for this device or not. If not, a new AL_PA is assigned to the NL_Port, which then causes the start of another LIP. This ensures that the switch assigned AL_PA does not conflict with any previously selected AL_PAs on the loop.

After the N_Port or public NL_Port gets its fabric address from FLOGI, it needs to register with the SNS. This is done with port login (PLOGI) at the address 0xFFFFFC. The device may register values for all or just some database objects, but the most useful are its 24-bit port address, 64-bit World Wide Port Name (WWPN), 64-bit World Wide Node Name (WWN), class of service parameters, FC-4 protocols supported, and port type, such as N_Port or NL_Port.

Private devices on NL_Ports

It is easy to explain how the port to World Wide Name address resolution works when a single device from an N_Port is connected to an F_Port, or when a public NL_Port device is connected to FL_Port in the switch. The SNS will add an entry for the device World Wide Name and connects that with the port address which is selected from the selection of free port addresses for that switch. Problems may arise when a private Fibre Channel device is attached to the switch. Private Fibre Channel devices were designed to only work in private loops.

When the arbitrated loop is connected to the FL_Port, this port obtains the highest priority address in the loop to which it is attached (0x00). Then the FL_Port performs a LIP. After this process is completed, the FL_Port registers all devices on the loop with the SNS. Devices on the arbitrated loop use only 8-bit addressing, but in the switched fabric, 24-bit addressing is used. When the FL_Port registers the devices on the loop to the SNS, it adds two most significant bytes to the existing 8-bit address.

The format of the address in the SNS table is 0xPPPPLL, where the PPPP is the two most significant bytes of the FL_Port address and the LL is the device ID on the arbitrated loop which is connected to this FL_Port. Modifying the private loop address in this fashion, all private devices can now talk to all public devices, and all public devices can talk to all private devices.

Because we have stated that private devices can only talk to devices with private addresses, some form of translation must take place. We show an example of this in Figure 2-17.

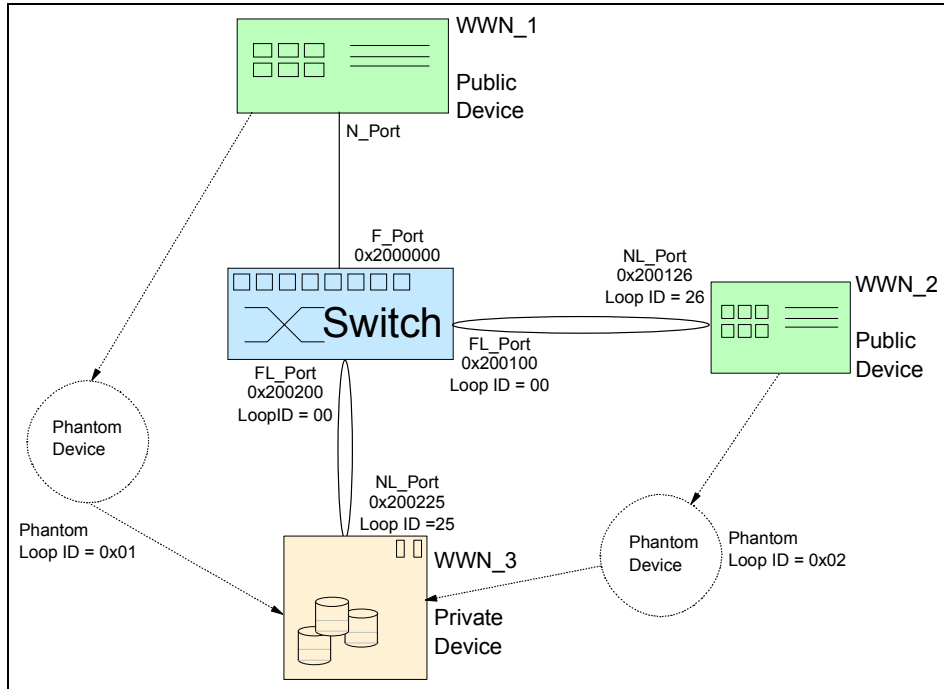


Figure 2-17 Arbitrated loop address translation

As you can see, we have three devices connected to the switch:

- ▶ Public device N_Port with WWN address WWN_1 on F_Port with the port address 0x200000.
- ▶ Public device NL_Port with WWN address WWN_2 on FL_Port with the port address 0x200100. The device has AL_PA 0x26 on the loop which is attached on the FL_Port
- ▶ Private device NL_Port with WWN address WWN_3 on FL_Port with the port address 0x200200. The device has AL_PA 0x25 on the loop which is attached to the FL_Port.

After all FLOGI and PLOGI functions are performed, the SNS will have the entries shown in Table 2-4.

Table 2-4 Name server entries

24-bit port address	WWN	FL_Port address
0x200000	WWN_1	n/a
0x200126	WWN_2	0x200100
0x200225	WWN_3	0x200200

We now explain some possible scenarios.

Public N_Port device accesses private NL_Port device

The communication from device to device starts with PLOGI to establish a session. When a public N_Port device wants to perform a PLOGI to a private NL_Port device, the FL_Port on which this private device exists will assign a “phantom” private address to the public device. This phantom address is known only inside this loop, and the switch keeps track of the assignments.

In our example, when the WWN_1 device wants to talk to the WWN_3 device, the addresses and IDs shown in Table 2-5 are created in the switch.

Table 2-5 Phantom addresses

Switch port address	Phantom loop port ID
0x200000	0x01
0x200126	0x02

When the WWN_1 device enters into the loop, it represents itself with AL_PA ID 0x01 (its phantom address). All private devices on that loop use this ID to talk to that public device. The switch itself acts as a proxy, and translates addresses in both directions. Usually the number of phantom addresses is limited, and this number of phantom addresses decreases the number of devices allowed in the arbitrated loop. For example, if the number of phantom addresses is 32, this limits the number of physical devices in the loop to $126 - 32 = 94$.

Public N_Port device accesses public NL_Port device

If an N_Port public device wants to access an NL_Port public device, it simply performs a PLOGI with the whole 24-bit address.

Private NL_Port device accesses public N_Port or NL_Port device

When a private device needs to access a remote public device, it uses the public device’s phantom address. When the FL_Port detects the use of a phantom AL_PA ID, it translates that to a switch port ID using its translation table similar to that shown in Table 2-5.

Translative mode

As explained above, private devices can cooperate in the fabric using translative mode. However, if you have a private host (server), this is not possible. To solve this problem, switch vendors, including IBM, support a translative feature. This feature, often referred to and defined in the FC standards as TL_Mode, allows the whole switch or director, or just a set of ports, to operate as an arbitrated loop. In this mode, devices connected to the switch do not perform a fabric login, and the switch itself will emulate the loop for those devices. All public devices can still see all private devices in translative mode. This is described comprehensively in “Private devices on NL_Ports” on page 45.

Switching mechanism and performance

In a switched fabric, a “cut-through” switching mechanism is used. This is not unique to switched fabrics and it is also used in Ethernet switches. The function is to speed packet routing from port to port.

When a frame enters the switch, cut-through logic examines only the link level destination ID of the frame. Based on the destination ID, a routing decision is made, and the frame is switched to the appropriate port by internal routing logic contained in the switch. It is this cut-through which increases performance by reducing the time required to make a routing decision. The reason for this is that the destination ID resides in the first four bytes of the frame header, and this allows the cut-through to be accomplished quickly. A routing decision can be made at the instant the frame enters the switch, without interpretation of anything other than the four bytes.

Switch frame buffering

An important criterion in selecting a switch is the number of frames that can be buffered on the port. During periods of high activity and frame movement, the switch may not be able to transmit a frame to its intended destination. This is true if two ports are sending data to the same destination. Given this situation, but depending on the class of service, the switch may sacrifice the frames it is not able to process. Not only does frame buffering reduce this likelihood, it also enhances performance.

Domain number routing decision

Another great performance improvement can be realized in the way in which the 24-bit port address is built. Because the address is divided into domain, area, and port, it is possible to make the routing decision on a single byte. As one example of this, if the domain number of the destination address indicates that the frame is intended for a different switch, the routing process can forward the frame to the appropriate interconnection without the need to process the entire 24-bit address and the associated overhead.

Data path in switched fabric

A complex switched fabric can be created by interconnecting Fibre Channel switches. Switch-to-switch connections are performed by E_Port connections. This means that if you want to interconnect switches, they need to support E_Ports. Switches may also support multiple E_Port connections to expand the bandwidth.

In such a configuration with interconnected switches, known as a meshed topology, multiple paths from one N_Port to another can exist.

An example of a meshed topology is shown in Figure 2-18.

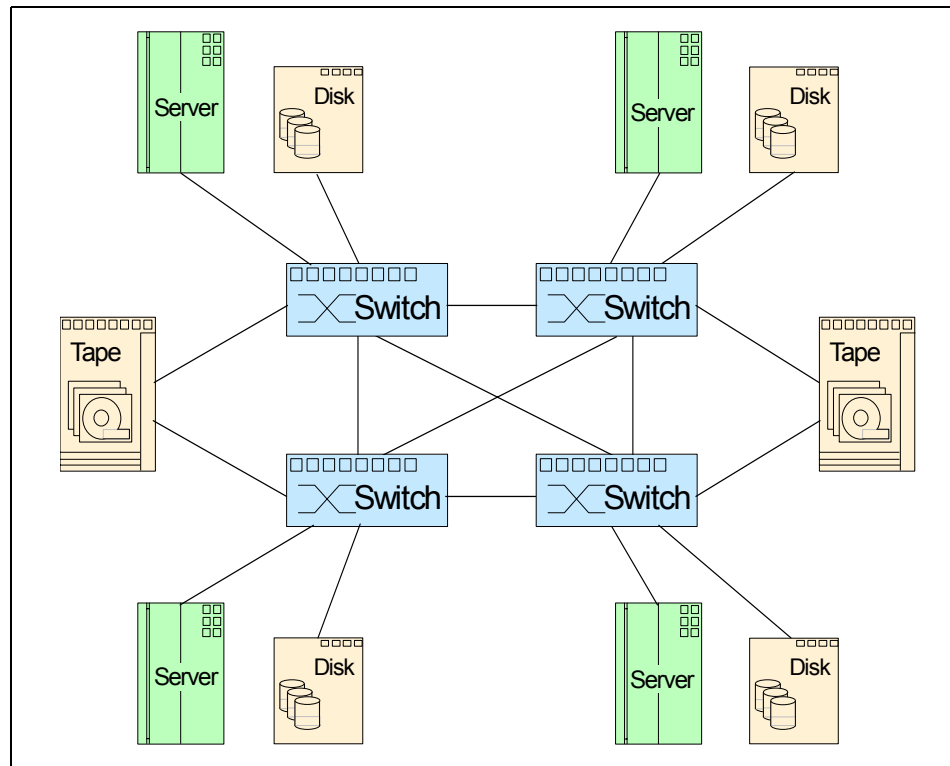


Figure 2-18 Meshed topology switched fabric

Spanning tree

In case of failure, it is important to consider having an alternative path between source and destination available. This will allow the data still to reach its destination. However, having different paths available could lead to the delivery of frames being out of the order of transmission, due to a frame taking a different path and arriving earlier than one of its predecessors.

A solution, which can be incorporated into the meshed fabric, is called a spanning tree and is an IEEE 802.1 standard. This means that switches keep to certain paths, as the spanning tree protocol will block certain paths to produce a simply connected active topology. Then the shortest path in terms of hops is used to deliver the frames and, most importantly, only one path is active at a time. This means that all associated frames go over the same path to the destination. The paths that are blocked can be held in reserve and used only if, for example, a primary path fails. The fact that one path is active at a time means that in the case of a meshed fabric, all frames will arrive in the expected order.

Path selection

For path selection, link state protocols are popular and extremely effective in today's networks. Examples of link state protocol are OSPF for IP and PNNI for ATM.

The most commonly used path selection protocol is Fabric Shortest Path First (FSPF). This type of path selection is usually performed at boot time, and no configuration is needed. All paths are established at start time, and only if the inter switch link (ISL) is broken or added will reconfiguration take place.

If multiple paths are available and if the primary path goes down, the traffic will be rerouted to another path. If the route fails, this can lead to congestion of frames, and any new frames delivered over the new path could potentially arrive at the destination first. This will cause an out-of-sequence delivery.

One possible solution for this situation is to prevent the activation of the new route for a while (this can be configured from milliseconds to a few seconds), so the congested frames are either delivered or rejected. Obviously, this can slow down the routing, so it should only be used when the devices connected to the fabric are not in a position to, or cannot tolerate, occasional out-of-sequence delivery. For instance, video can tolerate an out-of-sequence delivery, but financial and commercial data cannot.

But today, Fibre Channel devices are much more sophisticated, and this is a feature that is not normally required. FSPF allows a fabric to still benefit from load balancing the delivery of frames by using multiple paths.

We discuss FSPF in greater depth in “Fabric shortest path first” on page 91.

Route definition

Routes are usually dynamically defined. Static routes can also be defined. In the event that a static route fails, a dynamic route will take over. Once the static route becomes available again, frames will return to utilizing that route.

If dynamic paths are used, FSPF path selection is used. This guarantees that only the shortest and fastest paths will be used for delivering the frames.

We show an example of FSPF in Figure 2-19.

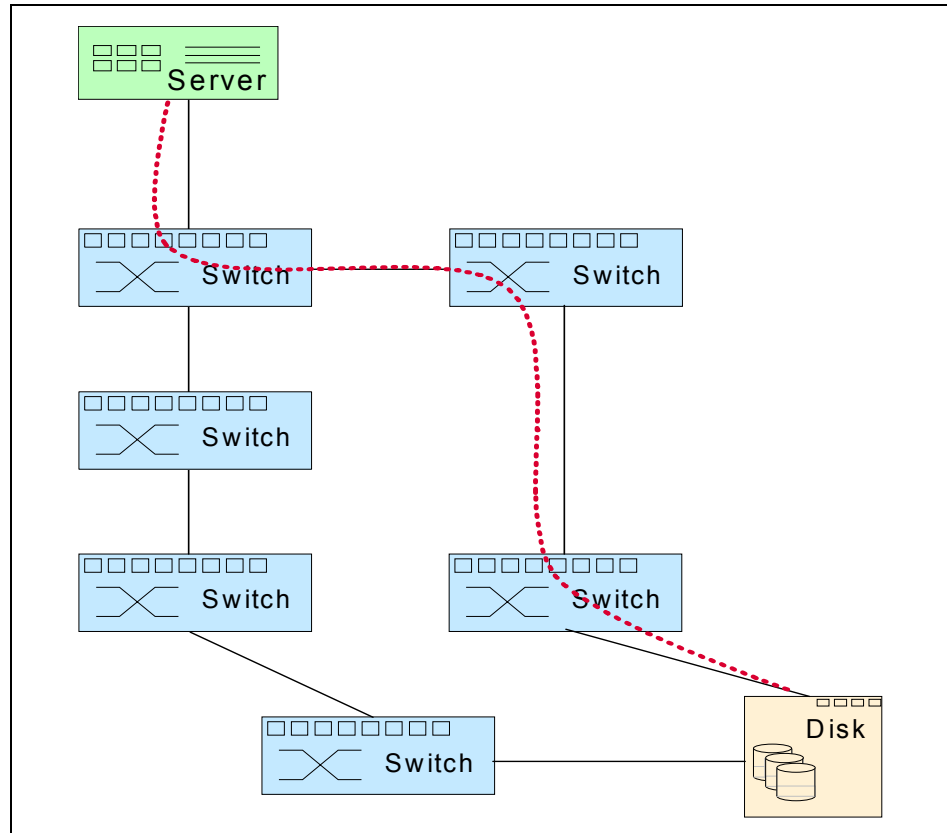


Figure 2-19 Fabric shortest path first

Adding new devices

Switched fabrics, by their very nature, are dynamic environments. They can handle topology changes as new devices are attached, or previously active devices are removed or taken offline. For these reasons it is important that notification of these types of events can be provided to participants (nodes) in the switched fabric.

Notification is provided by two functions:

- ▶ State Change Notification: SCN
- ▶ Registered State Change Notification: RSCN

These two functions are not obligatory, so each N_Port or NL_Port must register its interest in being notified of any topology changes, or if another device alters its state.

The original SCN service allowed an N_Port to send a notification change directly to another N_Port. This is not necessarily an optimum solution, as no other participants on the fabric will know about this change. RSCN offers a solution to this and will inform all registered devices about the change.

Perhaps the most important change that you would want to be notified about, is when an existing device goes offline. This information is very meaningful for participants that communicate with that device. For example, a server in the fabric environment would want to know if their resources are powered off or removed, or when new resources became available for its use.

Changed notification provides the same functionality for the switched fabric as loop initialization provides for arbitrated loop.

2.4.5 WWN and WWPN

Each device in the SAN is identified by a unique world wide name (WWN). The WWN contains a vendor identifier field, which is defined and maintained by the IEEE, and a vendor specific information field.

For further information, visit the Web site:

<http://standards.ieee.org/>

Currently, there are two formats of the WWN as defined by the IEEE. The original format contains either a hex 10 or hex 20 in the first two bytes of the address. This is then followed by the vendor specific information.

The new addressing scheme starts with a hex 5 or 6 in the first half-byte followed by the vendor identifier in the next 3 bytes. The vendor specific information is then contained in the following fields.

Both the old and new WWN formats are shown in Figure 2-20.

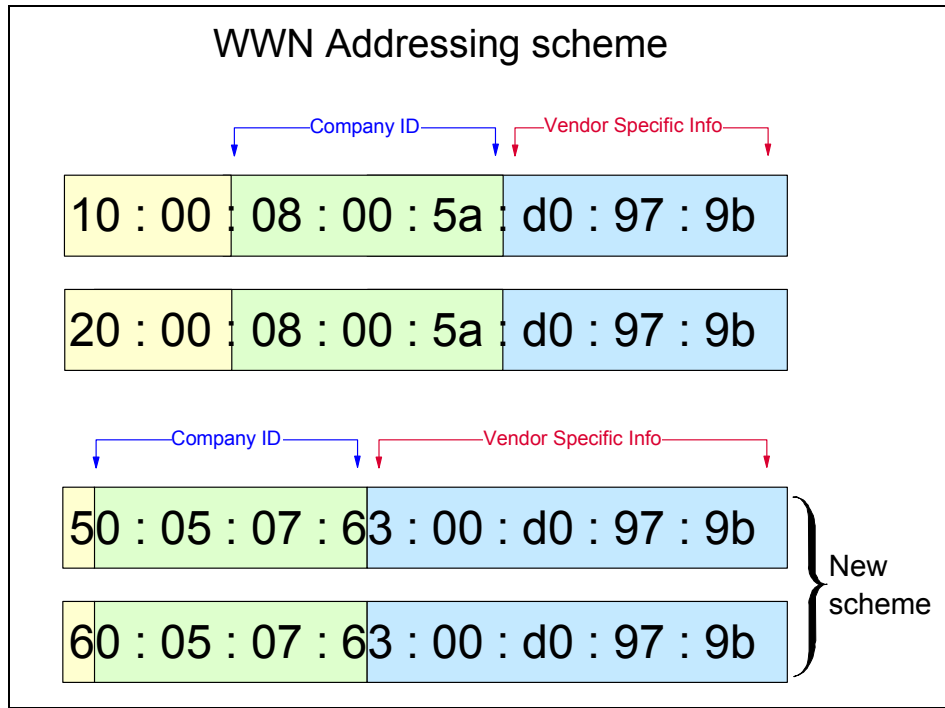


Figure 2-20 World Wide Name addressing scheme

The complete list of vendor identifiers as maintained by the IEEE is available at:

<http://standards.ieee.org/regauth/oui/oui.txt>

Some devices may have multiple Fibre Channel adapters, like an ESS, for example. In this case the device also has an identifier for each of its Fibre Channel adapters. This identifier is called the world wide port name (WWPN). This way it is possible to uniquely identify all Fibre Channel adapters and paths within a device.

This is illustrated in Figure 2-21.

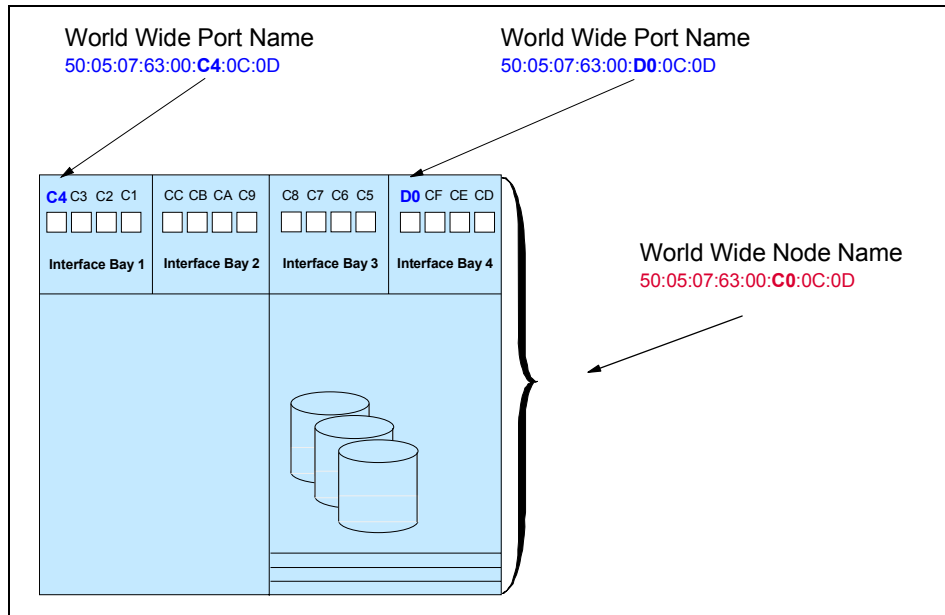


Figure 2-21 WWN and WWPN

This diagram shows how each of the ESS's Fibre Channel adapters has a unique WWPN. In the case of the ESS, the vendor specific information field is used to identify each Fibre Channel adapter according to which bay and slot position it is installed in within the ESS.

Shown in Figure 2-22 is a screen capture of the name server table for a test SAN in the ITSO lab. This shows that the two devices (DEC HSG80 and IBM 1742) both have multiple HBAs. The name server table shows the WWN for each device as being the same, but the WWPN is different for each HBA within these devices.

Name Server Table

☐ Auto Refresh Auto-Refresh Interval: Seconds

Domain #	Port #	Port ID	Port Type	Port WWN	Node WWN	Symbolic Name
3	3	031300	N	50:00:1fe1:00:0a:e4:02	50:00:1fe1:00:0a:e4:00	[28] "DEC HSG80 V
3	4	031400	N	20:02:00:a0:b8:0c:bc:e8	20:02:00:a0:b8:0c:bc:e7	[28] "IBM 1742 052
3	6	031600	N	50:00:1fe1:00:0a:e4:01	50:00:1fe1:00:0a:e4:00	[28] "DEC HSG80 V
3	7	031700	N	20:03:00:a0:b8:0c:bc:e8	20:02:00:a0:b8:0c:bc:e7	[28] "IBM 1742 052
3	10	031a00	N	21:00:00:e0:8b:05:f0:ed	20:00:00:e0:8b:05:f0:ed	NULL
3	11	031b00	N	20:04:00:60:45:16:0d:2e	10:00:00:60:45:16:0d:2e	[28] "PATHLIGHTSAN Gatew
3	14	031e00	N	21:00:00:e0:8b:05:49:bc	20:00:00:e0:8b:05:49:bc	NULL
3	15	031f00	N	21:00:00:e0:8b:05:46:bc	20:00:00:e0:8b:05:46:bc	NULL

Figure 2-22 WWN and WWPN entries in a name server table

2.4.6 Zoning

Zoning allows for finer segmentation of the switched fabric. Zoning can be used to instigate a barrier between different environments. Only members of the same zone can communicate within that zone, and all other attempts from outside the zone are rejected.

For example, it may be desirable to separate a Windows environment from a UNIX environment. This is very useful because of the manner in which Windows attempts to claim all available storage for itself. Because not all storage devices are capable of protecting their resources from any host searching for available resources, it makes sound business sense to protect the environment.

Looking at zoning in this way, it could also be considered as a security feature and not just for separating environments. Zoning could also be used for test and maintenance purposes. For example, not many enterprises will mix their test and maintenance environments with their production environment. Within a fabric, you could easily separate your test environment from your production bandwidth allocation on the same fabric using zoning.

In fact, it is of historical note that zoning was developed to prevent some operating systems from writing their signature on all devices that they saw. This would mean that unsuspecting operating systems that were less parochial in nature had the potential to lose access to their disks.

We show an example of zoning in Figure 2-23 where we have separated AIX® from NT and created Zone 1 and Zone 2. This diagram also shows how a device can be in more than one zone.

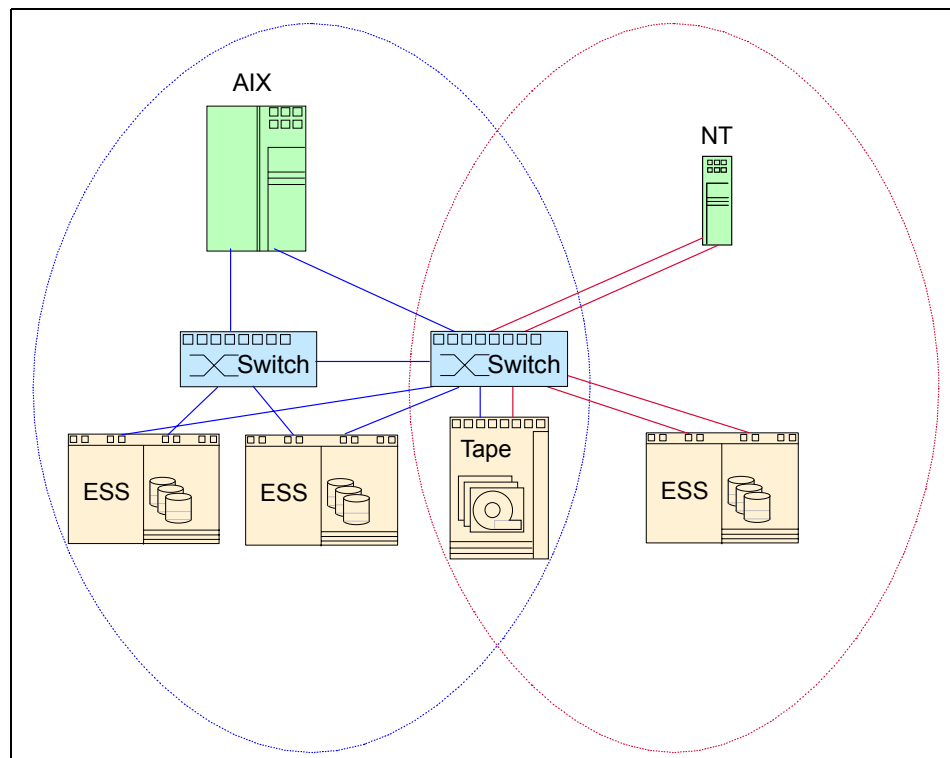


Figure 2-23 Zoning

Zoning also introduces the flexibility to manage a switched fabric to meet different user groups objectives.

Implementing zoning

Zoning can be implemented in two ways:

- ▶ Hardware zoning
- ▶ Software zoning

Hardware or port zoning

Hardware zoning is based on the physical fabric port number. The members of a zone are physical ports on the fabric switch. It can be implemented in the following configurations:

- ▶ One-to-one
- ▶ One-to-many
- ▶ Many-to-many

A single port can also belong to multiple zones. We show an example of hardware zoning in Figure 2-24.

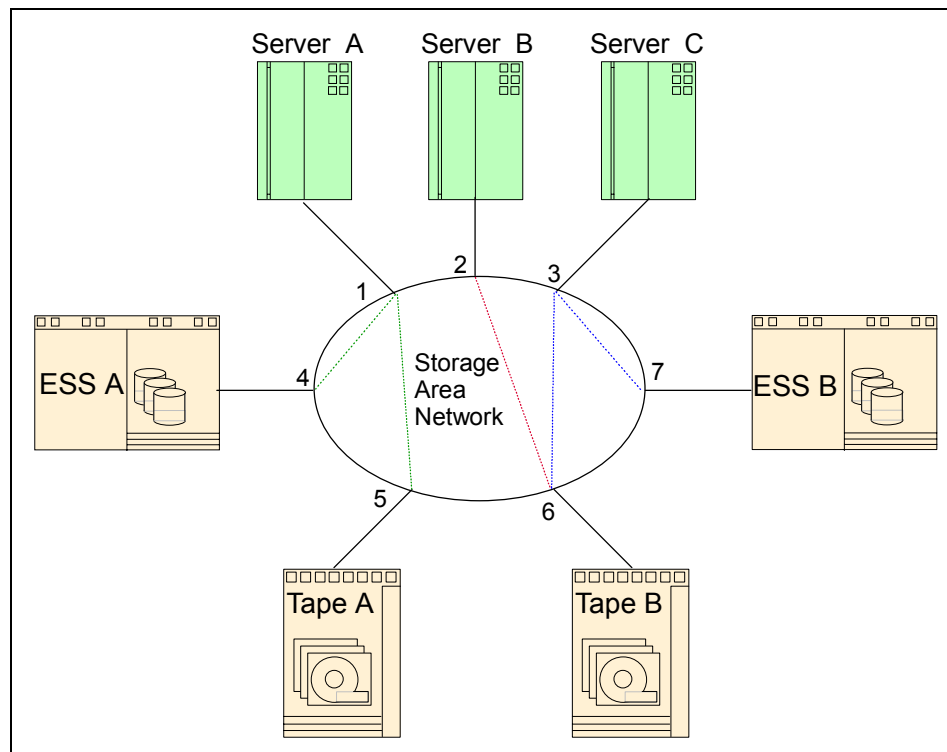


Figure 2-24 Zoning based on the switch port-number

In this example, port-based zoning is used to restrict Server A to only see storage devices that are zoned to port 1, that is, ports 4 and 5.

Server B is also zoned so that it can only see from port 2 through to port 6.

Server C is zoned so that it can see both ports 6 and 7 even though port 6 is also a member of another zone.

One of the disadvantages of port-based zoning is that devices have to be connected to a specific port, and the whole zoning configuration could become unusable if the device is connected to a different port.

For example, if the device attached to port 4 was removed and re-cabled into port 7, Server C would be able to see through to ESS A. That could cause an issue if that server is not allowed to see this device.

This example could also occur if port 4 failed requiring ESS A to be re-cabled into a new port (port 8 for example). Any zone containing reference to port 4 would need to be replaced with the new port number (port 8 in this case). This would involve having to manually update the zoning information and applying the change to the fabric.

The advantage of port-based zoning is that it can be implemented into a routing engine by filtering. As a result, this kind of zoning has a very low impact on the performance of the routing process.

In cases where the device connections are not permanent the use of WWN zoning is recommended.

Software or WWN zoning

Software zoning is implemented within the name server running inside the fabric switch. When using software zoning the members of the zone can be defined by:

- ▶ Node WWN
- ▶ Port WWN

Usually zoning software also allows you to create symbolic names or aliases for the zone members and for the zones themselves. Dealing with the symbolic name or aliases for a device is often easier than trying to use the WWN address, which, for example, is in the format of 20:0a:00:ab:cd:12:23:34.

The number of members possible in a zone is limited only by the amount of memory in the fabric switch. A member can belong to multiple zones. You can define multiple configurations or sets of zones for the fabric, but only one configuration or set can be active or enabled at any time. You can activate another zone set or configuration any time you want, without the need to power down the switch.

With software zoning there is no need to worry about the physical connections to the switch. If you use WWNs for the zone members, even when a device is connected to another physical port, it will still remain in the same zoning definition, because the device's WWN remains the same. The zone follows the WWN.

Shown in Figure 2-25 is an example of WWN based zoning. In this example symbolic names are defined for each WWN in the SAN to implement the same zoning requirements, as shown in the previous example for port zoning:

- ▶ **Zone_1** contains the aliases **alex**, **ben**, and **sam**, and is restricted to only these devices.
- ▶ **Zone_2** contains the aliases **robyn** and **ellen**, and is restricted to only these devices.
- ▶ **Zone_3** contains the aliases **matthew**, **max**, and **ellen**, and is restricted to only these devices.

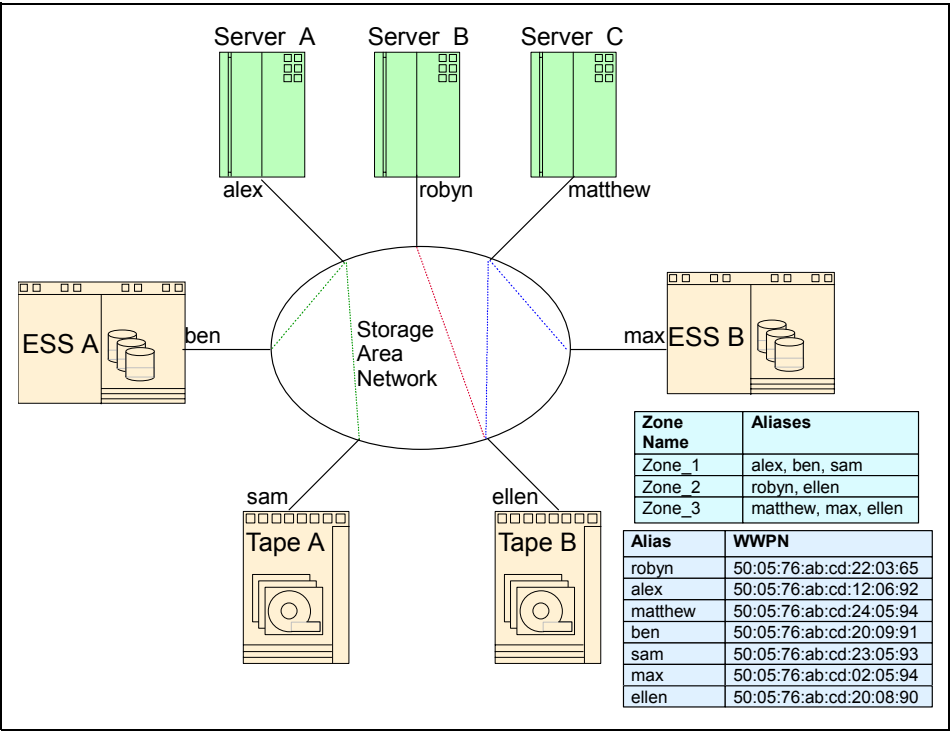


Figure 2-25 Zoning based on the devices WWN

There can be a potential security leak with software zoning. When a specific host logs into the fabric and asks for available storage devices, the name server will query the software zoning table to see which storage devices are allowable for that host. The host will only see the storage devices defined in the software zoning table. But the host can also make a direct connection to the storage device, while doing device discovery, without asking the name server for the information it has.

Additionally, any device that does any form of probing for WWNs may be able to discover devices and talk to them. A simple analogy might be that of an unlisted telephone number where, although the telephone number is not publicly available, there is nothing to stop a person from dialing that number whether by design or accident. The same holds true for WWNs, and there are devices that will randomly probe for WWNs to see if they can start a conversation with them. These are known as “bad citizens”.

A number of switch vendors offer hardware enforced WWN zoning, which can prevent this security exposure.

Note: For maximum security, hardware zoning is recommended. But as the standards are evolving and the industry is following them, it is likely that in the future, software zoning will probably be the preferred solution.

LUN masking

Another approach to securing storage devices from hosts wishing to take over already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts by means of LUNs.

For example, each partition in the storage server has its own LUN. If the host (server) wants to access the storage, it needs to request access to the LUN in the storage device. The purpose of LUN masking is to control access to the LUNs. The storage device itself accepts or rejects access requests from different hosts.

The user defines which hosts can access which LUN by means of the storage device control program. Whenever the host accesses a particular LUN, the storage device will check its access list for that LUN, and it will allow or disallow access to the LUN.

2.4.7 Expanding the fabric

As the demand for access to the storage grows, a switched fabric can be expanded to service these needs. Not all storage requirements can be satisfied with fabrics alone. For some applications, the 200 MB/s per port and advanced services are overkill, and they amount to wasted bandwidth and unnecessary cost. When you design a storage network, you need to consider the application's needs and not just rush to implement the latest technology available. SANs are often combinations of switched fabric and arbitrated loops.

Cascading

Expanding the fabric is called switch cascading. Cascading is basically interconnecting Fibre Channel switches and/or directors. The cascading of switches provides the following benefits to a SAN environment:

- ▶ The fabric can be seamlessly extended. Additional switches can be added to the fabric, without powering down existing fabric.
- ▶ You can easily increase the distance between various SAN participants.
- ▶ By adding more switches to the fabric, you increase connectivity by providing more available ports.
- ▶ Cascading provides high resilience in the fabric.
- ▶ With inter-switch links (ISLs), you can increase the bandwidth. The frames between the switches are delivered over all available data paths. So the more ISLs you create, the faster the frame delivery will be, but careful consideration must be employed to ensure that a bottleneck is not introduced.
- ▶ When the fabric grows, the name server is fully distributed across all the switches in fabric.
- ▶ With cascading, you also provide greater fault tolerance within the fabric.

Hops

As we stated in “Name and addressing” on page 42, the maximum number of switches allowed in the fabric is 239. The other limitation is that currently only seven hops are allowed between any source and destination, however, this is subject to change.

For the most up-to-date information refer to:

<http://www.storage.ibm.com/ibmsan/products/cisco/library.html>

We show a sample configuration that illustrates this in Figure 2-26, with “Hoppy”, the hop count kangaroo.

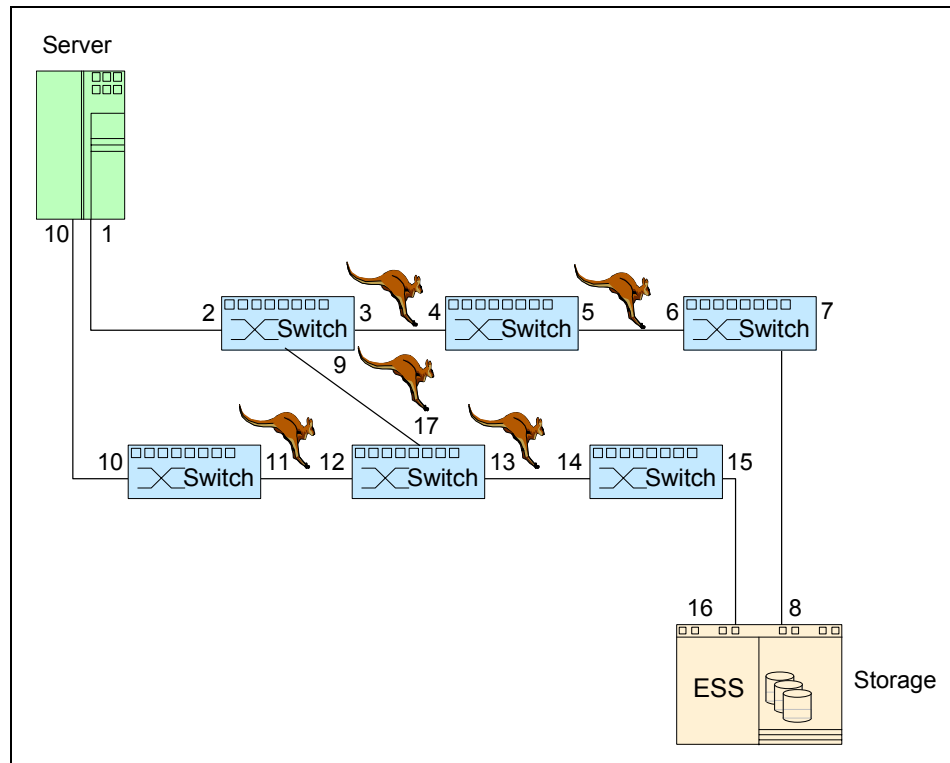


Figure 2-26 Cascading in a switched fabric

The hop count limit is set by the fabric operating system and is used to derive a frame holdtime value for each switch. This holdtime value is the maximum amount of time that a frame can be held in a switch before it is dropped (Class 3) or the fabric is busy (F_BSY, Class 2) is returned. A frame would be held if its destination port is not available. The holdtime is derived from a formula using the error detect time-out value (E_D_TOV) and the resource allocation time-out value (R_A_TOV).

The value of seven hops is not “hard-coded”, and if manipulation of E_D_TOV or R_A_TOV were to take place, the reasonable limit of seven hops could be exceeded. However, be aware that any hop suggestion was not a limit that was arrived at without careful consideration of a number of factors. In the future, the number of hops is likely to increase.

2.5 SAN software management standards

Traditionally, storage management has been the responsibility of the host server to which the storage resources are attached. With storage networks the focus has shifted away from individual server platforms, making storage management independent of the operating system, and offering the potential for greater flexibility by managing shared resources across the enterprise SAN infrastructure. Software is needed to configure, control, and monitor the SAN and all of its components in a consistent manner. Without good software tools, SANs cannot be implemented effectively.

The management challenges faced by SANs are very similar to those previously encountered by LANs and WANs. Single vendor proprietary management solutions will not satisfy customer requirements in a multi-vendor heterogeneous environment. The pressure is on the vendors to establish common methods and techniques. For instance, the need for platform independence for management applications, to enable them to port between a variety of server platforms, has encouraged the use of Java™.

The Storage Network Management Working Group (SNMWG) of SNIA is working to define and support open standards needed to address the increased management requirements imposed by SAN topologies. Reliable transport of the data, as well as management of the data and resources (such as file access, backup, and volume management) are key to stable operation. SAN management requires a hierarchy of functions, from management of individual devices and components, to the network fabric, storage resources, data, and applications. This is shown in Figure 2-27.

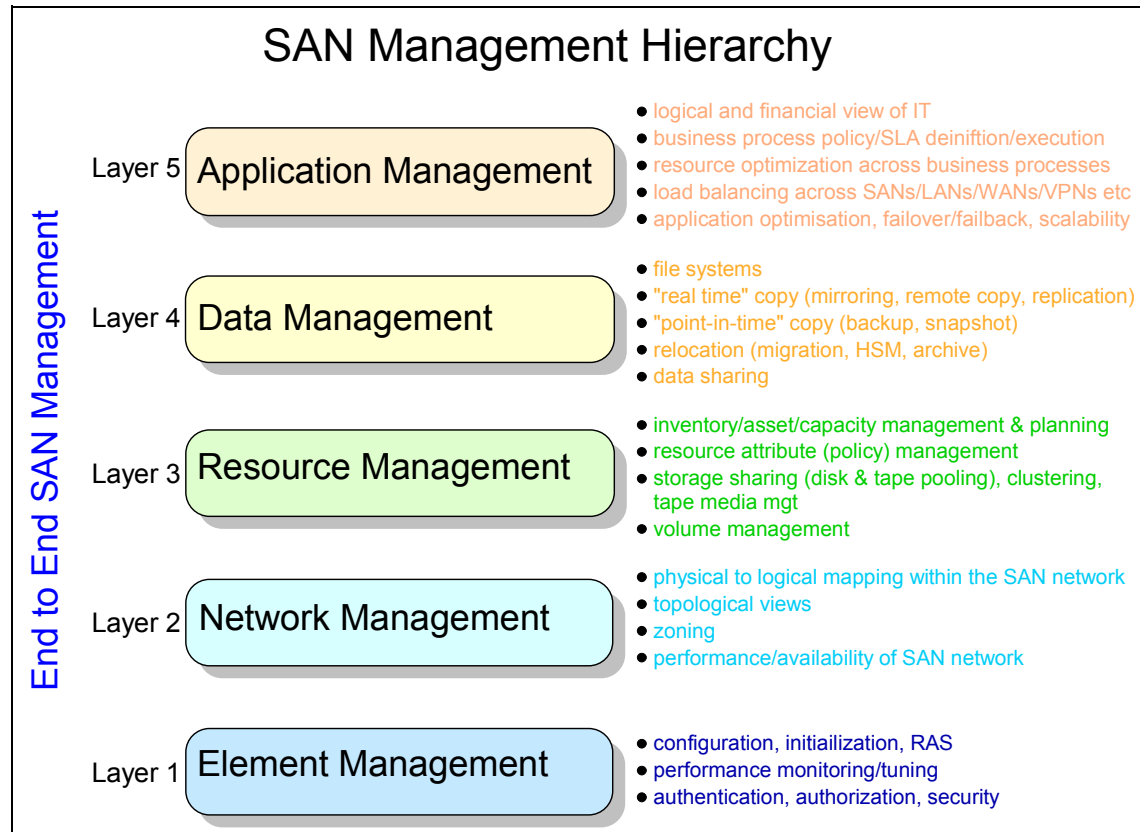


Figure 2-27 SAN management hierarchy

These can be implemented separately, or potentially, as a fully integrated solution to present a single interface to manage all SAN resources.

2.5.1 Application management

Application management is concerned with the availability, performance, and recoverability of the applications that run your business. Failures in individual components are of little consequence if the application is unaffected. By the same measure, a fully functional infrastructure is of little use if it is configured incorrectly or if the data placement makes the application unusable. Enterprise application and systems management is at the top of the hierarchy and provides a comprehensive, organization-wide view of all network resources (fabric, storage, servers, applications).

A flow of information regarding configuration, status, statistics, capacity utilization, performance, and so on, must be directed up the hierarchy from lower levels. A number of industry initiatives are directed at standardizing the storage specific information flow using a Common Information Model (CIM) or application programming interfaces (API), such as those proposed by the Jiro initiative, sponsored by Sun Microsystems, and others by SNIA and SNMWG.

Figure 2-28 illustrates a common interface model for heterogeneous, multi-vendor SAN management.

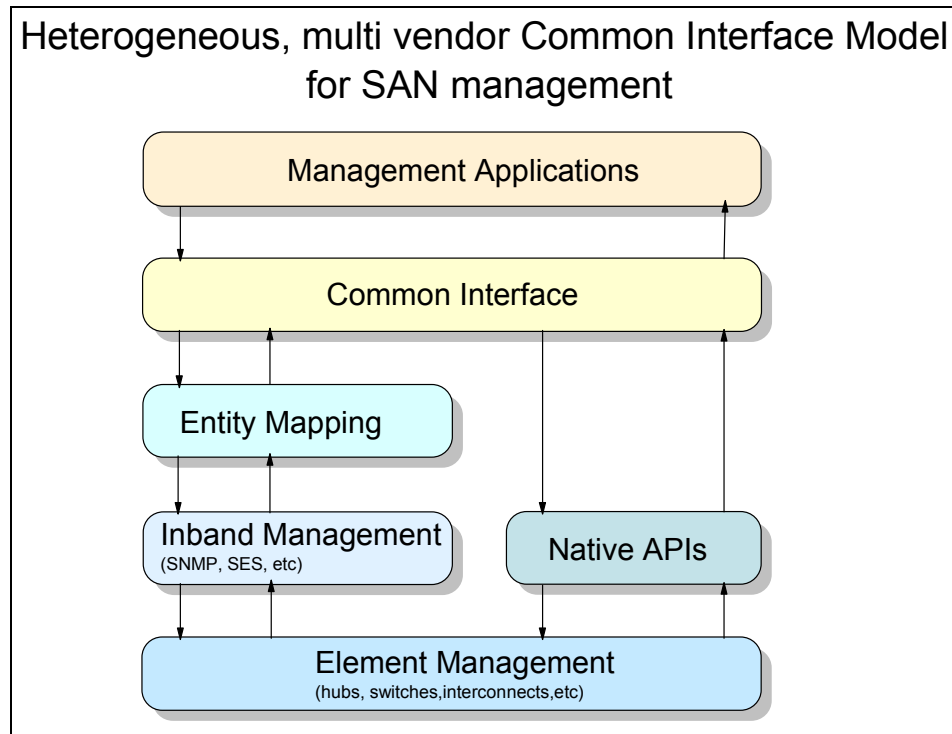


Figure 2-28 Common Interface Model for SAN management

2.5.2 Data management

More than at any other time in history, digital data is fueling business. Data management is concerned with Quality-of-Service (QoS) issues surrounding this data, such as:

- ▶ Ensuring data availability and accessibility for applications
- ▶ Ensuring proper performance of data for applications
- ▶ Ensuring recoverability of data

Data management is carried out on mobile and remote storage, centralized host attached storage, network attached storage (NAS), and SAN attached storage (SAS). It incorporates backup and recovery, archive and recall, and disaster protection.

2.5.3 Resource management

Resource management is concerned with the efficient utilization and consolidated, automated management of existing storage and fabric resources, as well as automating corrective actions where necessary. This requires the ability to manage all distributed storage resources, ideally through a single management console, to provide a single view of enterprise resources.

Without such a tool, storage administration is limited to individual servers. Typical enterprises today may have hundreds, or even thousands, of servers and storage subsystems. This makes impractical the manual consolidation of resource administration information, such as enterprise-wide disk utilization, or regarding the location of storage subsystems. SAN resource management addresses tasks such as:

- ▶ Pooling of disk resources
- ▶ Space management
- ▶ Pooling and sharing of removable media resources
- ▶ Implementation of “just-in-time” storage

2.5.4 Network management

Every e-business depends on existing LAN and WAN connections in order to function. Because of their importance, sophisticated network management software has evolved. Now SANs are allowing us to bring the same physical connectivity concepts to storage. And like LANs and WANs, SANs are vital to the operation of an e-business. Failures in the SAN can stop the operation of an enterprise.

SANs can be viewed as both physical and logical entities.

SAN physical view

The physical view identifies the installed SAN components, and allows the physical SAN topology to be understood. A SAN environment typically consists of four major classes of components:

- ▶ End-user computers and clients
- ▶ Servers
- ▶ Storage devices and subsystems
- ▶ Interconnect components

End-user platforms and server systems are usually connected to traditional LAN and WAN networks. In addition, some end-user systems may be attached to the Fibre Channel network, and may access SAN storage devices directly. Storage subsystems are connected using the Fibre Channel network to servers, end-user platforms, and to each other. The Fibre Channel network is made up of various interconnect components, such as switches, hubs, and bridges (Figure 2-29).

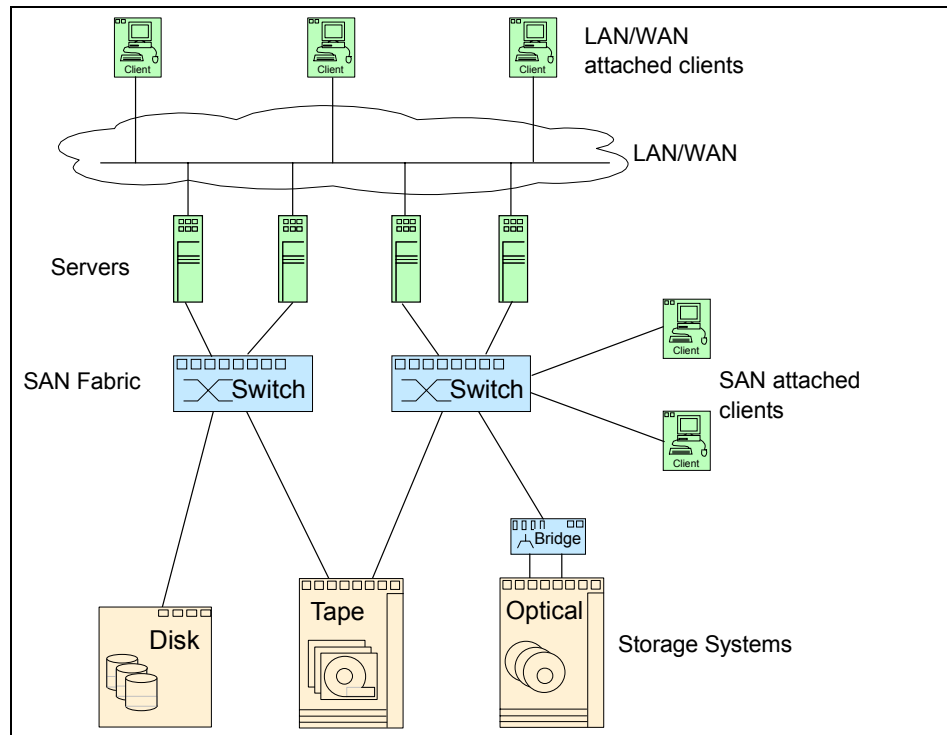


Figure 2-29 Typical SAN environment

SAN logical view

The logical view identifies and understands the relationships between SAN entities. These relationships are not necessarily constrained by physical connectivity, and they play a fundamental role in the management of SANs. For instance, a server and some storage devices may be classified as a logical entity. A logical entity group forms a private virtual network, or zone, within the SAN environment with a specific set of connected members. Communication within each zone is restricted to its members.

Network management is concerned with the efficient management of the Fibre Channel SAN — especially in physical connectivity mapping, fabric zoning, performance monitoring, error monitoring, and predictive capacity planning.

2.5.5 Element management

The elements that make up the SAN infrastructure include intelligent disk subsystems, intelligent removable media subsystems, Fibre Channel switches, hubs and bridges, meta-data controllers, and out-board storage management controllers. The vendors of these components provide proprietary software tools to manage their individual elements, usually comprising software, firmware, and hardware elements, such as those shown in Figure 2-30.

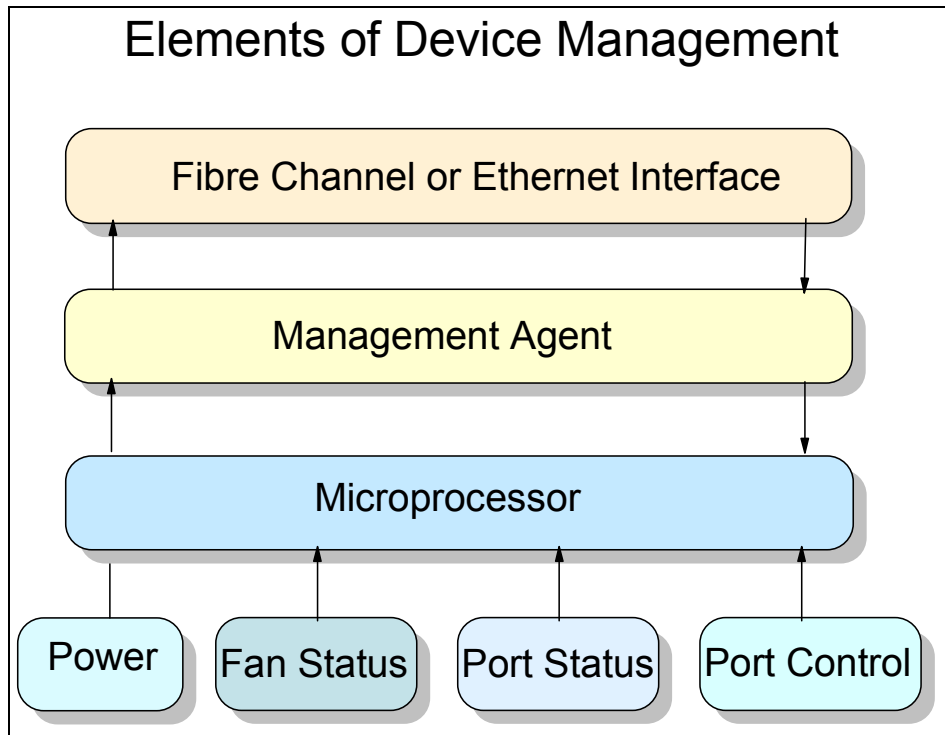


Figure 2-30 Device management elements

For instance, a management tool for a hub will provide information regarding its own configuration, status, and ports, but will not support other fabric components such as other hubs, switches, HBAs, and so on. Vendors that sell more than one element commonly provide a software package that consolidates the management and configuration of all of their elements. Modern enterprises, however, often purchase storage hardware from a number of different vendors.

Fabric monitoring and management is an area where a great deal of standards work is being focused. Two management techniques are in use — in-band and out-of-band management.

In-band management

Device communications to the network management facility is most commonly done directly across the Fibre Channel transport, using a protocol called SCSI Enclosure Services (SES). This is known as in-band management. It is simple to implement, requires no LAN connections, and has inherent advantages, such as the ability for a switch to initiate a SAN topology map by means of SES queries to other fabric components. However, in the event of a failure of the Fibre Channel transport itself, the management information cannot be transmitted. Therefore, access to devices is lost, as is the ability to detect, isolate, and recover from network problems. This problem can be minimized by provision of redundant paths between devices in the fabric.

- **In-band developments:** In-band management is evolving rapidly. Proposals exist for low level interfaces such as Return Node Identification (RNID) and Return Topology Identification (RTIN) to gather individual device and connection information, and for a Management Server that derives topology information. In-band management also allows attribute inquiries on storage devices and configuration changes for all elements of the SAN. Since in-band management is performed over the SAN itself, administrators are not required to make additional TCP/IP connections.

Out-of-band management

Out-of-band management means that device management data are gathered over a TCP/IP connection such as Ethernet. Commands and queries can be sent using Simple Network Management Protocol (SNMP), Telnet (a text-only command line interface), or a Web browser Hyper Text Transfer Protocol (HTTP). Telnet and HTTP implementations are more suited to small networks.

Out-of-band management does not rely on the Fibre Channel network. Its main advantage is that management commands and messages can be sent even if a loop or fabric link fails. Integrated SAN management facilities are more easily implemented, especially by using SNMP. However, unlike in-band management, it cannot automatically provide SAN topology mapping.

- **Out-of-band developments:** Two primary SNMP MIBs are being implemented for SAN fabric elements that allow out-of-band monitoring. The ANSI Fibre Channel Fabric Element MIB provides significant operational and configuration information on individual devices. The emerging Fibre Channel Management MIB provides additional link table and switch zoning information that can be used to derive information about the physical and logical connections between individual devices. Even with these two MIBs, out-of-band monitoring is incomplete. Most storage devices and some fabric devices don't support out-of-band monitoring. In addition, many administrators simply don't attach their SAN elements to the TCP/IP network.

- ▶ **Simple Network Management Protocol (SNMP):** This protocol is widely supported by LAN/WAN routers, gateways, hubs and switches, and is the predominant protocol used for multi vendor networks. Device status information (vendor, machine serial number, port type and status, traffic, errors, and so on) can be provided to an enterprise SNMP manager. This usually runs on a UNIX or NT workstation attached to the network. A device can generate an alert by SNMP, in the event of an error condition. The device symbol, or icon, displayed on the SNMP manager console, can be made to turn red or yellow, and messages can be sent to the network operator.
- ▶ **Management Information Base (MIB):** A management information base (MIB) organizes the statistics provided. The MIB runs on the SNMP management workstation, and also on the managed device. A number of industry standard MIBs have been defined for the LAN/WAN environment. Special MIBs for SANs are being built by the SNIA. When these are defined and adopted, multi-vendor SANs can be managed by common commands and queries.

Element management is concerned with providing a framework to centralize and automate the management of heterogeneous elements and to align this management with application or business policy.

2.5.6 Storage Management Initiative

The Storage Networking Industry Association (SNIA) has launched the Storage Management Initiative (SMI) to enable a standard to be developed that would provide a highly-functional open interface for the management of storage networks.

The goal of SMI is to produce a design specification that is based on the Common Information Model (CIM) and Web Based Enterprise Management (WBEM) standards.

Bluefin is the code name for a SAN Management specification that was developed by a group consisting of 16 SNIA members as a foundation for unifying the storage management industry on a management interface standard.

The Bluefin technology was developed with technical contributions from IBM and employs the Common Information Model (CIM) and Web Based Enterprise Management (WBEM) technology to discover and manage resources in a multi-vendor SAN through common interfaces. When implemented in products, the Bluefin technology will improve the usefulness of storage management applications and provide management interoperability in heterogeneous SANs. Bluefin has been presented as a technology contribution to SNIA.

IBM will incorporate the Bluefin specification into the Storage Tank™ SAN-wide file system and storage virtualization engine to extend the concept of SAN interoperability beyond basic system identification and monitoring to more comprehensive and efficient management capabilities. The Bluefin technology will also be incorporated into Tivoli Storage management offerings in the near future.

IBM's current and future storage software products will support interoperability in heterogeneous SANs by rapidly integrating and implementing Bluefin and other evolving standards for management of storage systems which are based on CIM.

An example of the Bluefin technology is the recent announcement between IBM and HP to cross license storage Application Programming Interfaces (API) to simplify storage management of both companies storage devices.

The cross licensing agreement will enable IBM software to use APIs to manage the HP Storage Works EMA (HSG80) and EVA (HSV110) arrays as well as the HP Storage Works MA8000 products. The APIs will also enable HP to manage the IBM Enterprise Storage Server through HP OpenView.

2.5.7 InfiniBand

This is a serial technology that can be implemented across either optical fiber or copper cabling. The parallel bus architecture has an inherent latency as it needs to wait for all the bits sent across a parallel link to arrive before it can send more data which slows a system down. InfiniBand is based around a serial link which reduces the number of pins and electrical interconnects that are required thereby reducing manufacturing costs and improving the reliability.

The InfiniBand Architecture is designed around a point-to-point, switched I/O fabric in which the devices are interconnected by cascaded switches. InfiniBand supports a range of applications and can be used to provide the backplane interconnect for a single host, or to building a complex system or network consisting of multiple independent and clustered hosts and I/O components.

An InfiniBand switch fabric looks quite similar to current Fibre Channel SANs. In this architecture, InfiniBand nodes (storage devices and servers) interconnect with one another over the InfiniBand I/O fabric.

For more information on the InfiniBand architecture, visit:

<http://www.infinibandta.org/ibta/>

2.6 Fabric management methods

The SAN fabric can be managed using several remote and local access methods. Each vendor will decide on the most appropriate methods to employ on their particular product. Not all vendors are the same and from a management point of view it makes sense to investigate the possibilities before any investment is made.

2.6.1 Common methods

There are several access methods for managing a switch or director. This is summarized in Table 2-6.

Switches can be accessed simultaneously from different connections. If this happens, changes from one connection may not be updated to the other, and some may be lost. Make sure when connecting with simultaneous multiple connections, that you do not overwrite the work of another connection.

Table 2-6 Comparison of management access methods

Management method	Description	Local	In-band (Fibre Channel)	Out-of-band (Ethernet)
Serial Port	CLI locally from serial port on the switch	Yes	No	No
Telnet	CLI remotely via Telnet	No	Yes	Yes
SNMP	Manage remotely using the simple network management protocol (SNMP)	No	Yes	Yes
Management Server	Manage with the management server	No	Yes	No
SES	Manage through SCSI-3 enclosure services	No	Yes	No
Web Tools	Manage remotely through graphical user interface	No	Yes	Yes

2.6.2 Hardware setup for switch management

To enable remote connection to the switch, the switch must have a valid IP address. Two IP addresses can be set; one for the external out-of-band Ethernet port and one for in-band Fibre Channel network access.

2.6.3 Managing with Telnet

To make a successful Telnet connection to a switch, the user needs:

- ▶ Switch name or IP address
- ▶ Username
- ▶ Password

Any host system that supports Telnet can be used to connect to the switch over the Ethernet. If the host supports a name server, the switch name can be used to effect the Telnet connection. If name service is not used to register network devices, then the IP address is used to connect to the switch. For example:

```
telnet [switch_name]
telnet 192.168.64.9
```

When the Telnet connection is made, the user is prompted for a user name and password. The following section defines the default user names and passwords supplied with the switch. Both of these can be changed by the switch administrator.

2.7 SAN standards

Given the strong drive towards SANs from users and vendors alike, one of the most critical success factors is the ability of systems and software from different vendors to operate together in a seamless way. Standards are the basis for the interoperability of devices and software from different vendors.

A good benchmark is the level of standardization in today's LAN and WAN networks. Standard interfaces for interoperability and management have been developed, and many vendors compete with products based on the implementation of these standards. Customers are free to mix and match components from multiple vendors to form a LAN or WAN solution. They are also free to choose from several different network management software vendors to manage their heterogeneous network.

The major vendors in the SAN industry recognize the need for standards, especially in the areas of interoperability interfaces and application programming interfaces (APIs), as these are the basis for wide acceptance of SANs. Standards will allow customers a greater breadth of choice, and will lead to the deployment of cross-platform, multi-vendor, enterprise-wide SAN solutions.

2.7.1 SAN industry associations and organizations

A number of industry associations, standards bodies, and company groupings are involved in developing, and publishing SAN standards. The major groups linked with SAN standards are shown in Figure 2-31.

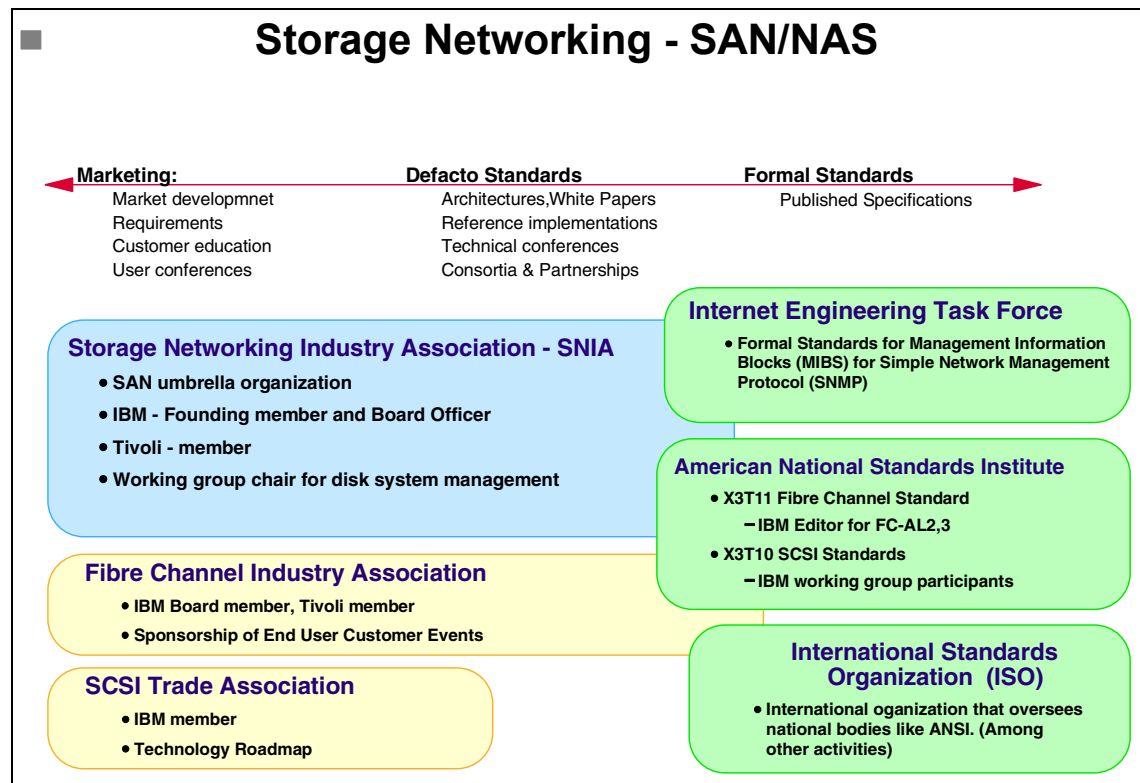


Figure 2-31 Groups involved in setting storage management standards

The roles of these associations and bodies fall into three categories:

- ▶ **Market development:** These associations are involved in market development, establishing requirements, conducting customer education, user conferences, and so on. The main organizations are the Storage Network Industry Association (SNIA); Fibre Channel Industry Association (merging the former Fibre Channel Association and the Fibre Channel Loop Community); and the SCSI Trade Association (SCSITA). Some of these organizations are also involved in the definition of defacto standards.
- ▶ **Defacto standards:** These organizations and bodies tend to be formed from two sources. They include working groups within the market development organizations, such as SNIA and FCIA. Others are partnerships between groups of companies in the industry, such as Jiro, Fibre Alliance, and the Open Standards Fabric Initiative (OSFI), which work as pressure groups towards defacto industry standards. They offer architectural definitions, write white papers, arrange technical conferences, and may reference implementations based on developments by their own partner companies. They may submit these specifications for formal standards acceptance and approval. The OSFI is a good example, comprising the five manufacturers of Fibre Channel switching products. In July 1999, they announced an initiative to accelerate the definition, finalization, and adoption of specific Fibre Channel standards that address switch interoperability.
- ▶ **Formal standards:** These are the formal standards organizations, like IETF, ANSI, and ISO, which are in place to review, obtain consensus, approve, and publish standards defined and submitted by the preceding two categories of organizations.

IBM and Tivoli Systems are heavily involved in most of these organizations, holding positions on boards of directors and technical councils and chairing projects in many key areas. We do this because it makes us aware of new work and emerging standards. The hardware and software management solutions we develop, therefore, can provide early and robust support for those standards that do emerge from the industry organizations into pervasive use. Secondly, IBM, as the innovation and technology leader in the storage industry, wants to drive reliability, availability, serviceability, and other functional features into standards. Following are the standards organizations in which we participate.

American National Standards Institute

American National Standards Institute (ANSI) does not itself develop American national standards. It facilitates development by establishing consensus among qualified groups. IBM participates in numerous committees, including those for Fibre Channel and storage area networks. For more information on ANSI, see its Web site at:

<http://www.ansi.org/>

INCITS

The International Committee for Information Technology Standards (INCITS) is the primary U.S. focus of standardization in the field of Information and Communications Technologies (ICT), encompassing storage, processing, transfer, display, management, organization, and retrieval of information. As such, INCITS also serves as ANSI's Technical Advisory Group for ISO/IEC Joint Technical Committee (JTC) 1. JTC 1 is responsible for International standardization in the field of Information Technology. From 1997 until 2001, INCITS operated under the name, Accredited Standards Committee, National Committee for Information Technology (NCITS). You'll find all their projects listed in 2.7.2, "List of evolved Fibre Channel standards" on page 79. For more information, see INCITS Web site at:

<http://www.incits.org/>

INCITS technical committee T11

Technical committee T11 retains overall responsibility for work in the area of "Device Level Interfaces" and does the proposals for Fibre Channel transport, Topology, Generic Services, and physical and media standards. The INCITS T11 committee is often referred to as the ANSI T11 group for short. Access to all proposals is available via the Web site at:

<http://www.t11.org>

Storage Networking Industry Association

Storage Networking Industry Association (SNIA) is an international computer industry forum of developers, integrators, and IT professionals who evolve and promote storage networking technology and solutions. SNIA was formed to ensure that storage networks become efficient, complete, and trusted solutions across the IT community. SNIA is accepted as the primary organization for the development of SAN standards, with over 125 companies as its members, including all the major server, storage, and fabric component vendors. SNIA also has a working group dedicated to the development of NAS standards, and is committed to delivering architectures, education, and services that will propel storage networking solutions into a broader market. IBM is one of the founding members of SNIA, and has senior representatives participating on the board and in technical groups. For additional information on the various activities of SNIA, see its Web site at:

<http://www.snia.org/home>

Fibre Channel Industry Association

The Fibre Channel Industry Association (FCIA) was formed in the autumn of 1999 as a result of a merger between the Fibre Channel Association (FCA) and the Fibre Channel Community (FCC). The FCIA currently has more than 150 members in the United States and through its affiliate organizations in Europe and Japan. The FCIA mission is to nurture and help develop the broadest market for Fibre Channel products. This is done through market development, education, standards monitoring, and fostering interoperability among members' products. IBM is a principal member of the FCIA.

Recently announced was the SANmark Qualified Program. The purpose of the program is to provide the industry with an objective indication of how Fibre Channel products perform against reasonable standards and to permit the use of the trademarked term *SANmark*, and any associated logo(s), in the identification and promotion of products meeting the published test indices.

The SANmark Qualified Program goals are to:

- ▶ Make Fibre Channel solutions easy to use, easy to install, manage, configure, diagnose, and troubleshoot
- ▶ Ensure that Fibre Channel continues to attain the highest performance and installed base maturity available in the market
- ▶ Proliferate heterogeneous shared SAN resources and heterogeneous management framework over WAN connections

For additional information on the various activities of FCIA, see its Web site at:

<http://www.fibrechannel.org/>

The SCSI Trade Association

The SCSI Trade Association (SCSITA) was formed to promote the use and understanding of small computer system interface (SCSI) parallel interface technology. The SCSITA provides a focal point for communicating SCSI benefits to the market, and influences the evolution of SCSI into the future. IBM is a founding member of the SCSITA. For more information, see its Web site at:

<http://www.scsita.org/>

InfiniBand (SM) Trade Association

The demands of the Internet and distributed computing are challenging the scalability, reliability, availability, and performance of servers. To meet this demand, a balanced system architecture with equally good performance in the memory, processor, and input/output (I/O) subsystems is required. A number of leading companies have joined together to develop a new common I/O specification beyond the current PCI bus architecture, to deliver a channel based, switched fabric technology that the entire industry can adopt. InfiniBand™ Architecture represents a new approach to I/O technology and is based on the collective research, knowledge, and experience of the industry's leaders. IBM is a founding member of InfiniBand (SM) Trade Association. For additional information, see its Web site at:

<http://www.infinibandta.org/home>

National Storage Industry Consortium

The National Storage Industry Consortium membership consists of over fifty US corporations, universities, and national laboratories with common interests in the field of digital information storage. A number of projects are sponsored by NSIC, including network attached storage devices (NASD), and network attached secure disks. The objective of the NASD project is to develop, explore, validate, and document the technologies required to enable the deployment and adoption of network attached devices, subsystems, and systems. IBM is a founding member of the NSIC. For more information, see its Web site at:

<http://www.nsic.org/>

Internet Engineering Task Force

The Internet Engineering Task Force (IETF) is a large, open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture, and the smooth operation of the Internet. It is responsible for the formal standards for the Management Information Blocks (MIB) and for Simple Network Management Protocol (SNMP) for SAN management. For additional information on IETF, see its Web site at:

<http://www.ietf.org/>

The IEEE Standards Association

The Institute of Electrical and Electronics Engineers Standards Association (IEEE-SA) is a membership organization that produces standards which are developed and used internationally, serving today's industries with a complete portfolio of standards programs. For more information on the IEEE-SA, see its Web site at:

<http://standards.ieee.org/sa/sa-view.html>

Distributed Management Task Force

The DMTF is the industry organization that is leading the development, adoption, and unification of management standards and initiatives for desktop, enterprise, and Internet environments. Working with key technology vendors and affiliated standards groups, the DMTF is enabling a more integrated, cost-effective, and less crisis-driven approach to management through interoperable management solutions. DMTF has released its WBEM Specifications (CIM Operations over HTTP v1.1, Representation of CIM in XML v2.1, and the CIM DTD v2.1.1) in final status. You can visit the Web site at:

<http://www.dmtf.org>

2.7.2 List of evolved Fibre Channel standards

Table 2-7 lists all current T11 Fibre Channel projects that are either approved standards or in proposal status. For the most recent status, visit the T11 Web site at:

<http://www.t11.org>

Table 2-7 T11 projects

Acronym	Title	Status
10 Bit Interface TR	10-bit Interface Technical Report	X3.TR-18:1997
10GFC	Fibre Channel - 10 Gigabit	Project 1413-D
FC-10KCR	Fibre Channel - 10 km Cost-Reduced Physical variant	INCITS 326: 1999
FC-AE	Fibre Channel Avionics Environment	INCITS TR-31-2002
FC-AE-2	Fibre Channel - Avionics Environment – 2	Project 1605-DT
FC-AL	FC Arbitrated Loop	ANSI X3.272:1996
FC-AL-2	Fibre Channel 2nd Generation Arbitrated Loop	INCITS 332: 1999
FC-AV	Fibre Channel - Audio-Visual	ANSI/INCITS 356:2001
FC-BB	Fibre Channel – Backbone	ANSI NCITS 342

Acronym	Title	Status
FC-BB-2	Fibre Channel - Backbone – 2	Project 1466-D
FC-CU	Fibre Channel Copper Interface Implementation Practice Guide	Project 1135-DT
FC-DA	Fibre Channel - Device Attach	Project 1513-DT
FC-FG	FC Fabric Generic Requirements	ANSI X3.289:1996
FC-FLA	Fibre Channel - Fabric Loop Attachment	INCITS TR-20:1998
FC-FP	FC - Mapping to HIPPI-FP	ANSI X3.254:1994
FC-FS	Fibre Channel Framing and Signaling Interface	Project 1331-D
FC-FS-2	Fibre Channel - Framing and Signaling – 2	Project
FC-GS	FC Generic Services	ANSI X3.288:1996
FC-GS-2	Fibre Channel 2nd Generation Generic Services	ANSI INCITS 288
FC-GS-3	Fibre Channel - Generic Services 3	NCITS 348-2000
FC-GS-4	Fibre Channel Generic Services 4	Project 1505-D
FC-HBA	Fibre Channel - HBA API	Project 1568-D
FC-HSPI	Fibre Channel High Speed Parallel Interface (FC-HSPI)	INCITS TR-26: 2000
FC-LE	FC Link Encapsulation	ANSI X3.287:1996
FC-LS	Fibre Channel - Link Services	Project

Acronym	Title	Status
FC-MI	Fibre Channel - Methodologies for Interconnects Technical Report	INCITS TR-30-2002
FC-MI-2	Fibre Channel - Methodologies for Interconnects – 2	Project 1599-DT
FC-MJS	Methodology of Jitter Specification	INCITS TR-25:1999
FC-MJSQ	Fibre Channel - Methodologies for Jitter and Signal Quality Specification	Project 1316-DT
FC-PH	Fibre Channel Physical and Signaling Interface	ANSI X3.230:1994
FC-PH-2	Fibre Channel 2nd Generation Physical Interface	ANSI X3.297:1997
FC-PH-3	Fibre Channel 3rd Generation Physical Interface	ANSI X3.303:1998
FC-PH:AM 1	FC-PH Amendment #1	ANSI X3.230:1994/AM1:1996
FC-PH:DAM 2	FC-PH Amendment #2	ANSI X3.230/AM2-1999
FC-PI	Fibre Channel - Physical Interface	INCITS 352
FC-PI-2	Fibre Channel - Physical Interfaces – 2	Project
FC-PLDA	Fibre Channel Private Loop Direct Attach	INCITS TR-19:1998
FC-SB	FC Mapping of Single Byte Command Code Sets	ANSI X3.271:1996
FC-SB-2	Fibre Channel - SB 2	INCITS 349-2000
FC-SB-3	Fibre Channel - Single Byte Command Set – 3	Project 1569-D

Acronym	Title	Status
FC-SP	Fibre Channel - Security Protocols	Project 1570-D
FC-SW	FC Switch Fabric and Switch Control Requirements	INCITS 321:1998
FC-SW-2	Fibre Channel - Switch Fabric – 2	ANSI/INCITS 355-2001
FC-SW-3	Fibre Channel - Switch Fabric – 3	Project 1508-D
FC-SWAPI	Fibre Channel Switch Application Programming Interface	Project 1600-D
FC-Tape	Fibre Channel - Tape Technical Report	INCITS TR-24:1999
FC-VI	Fibre Channel - Virtual Interface Architecture Mapping	ANSI/INCITS 357-2001
FCSM	Fibre Channel Signal Modeling	Project 1507-DT
MIB-FA	Fibre Channel Management Information Base	Project 1571-DT
SM-LL-V	FC - Very Long Length Optical Interface	ANSI/INCITS 339-2000
SM-AMD	SAN Management - Attribute & Method Dictionary	Project 1606-DT
SM-MM	SAN Management - Management Model	Project 1606-DT

10 Gb/s

10GFC is a working draft for the extensions to the FC-PH and FC-PI standard to support a data rate of 10.2 Gb/s. The proposal includes five different physical interface types — three shortwave and two longwave solutions:

- ▶ SW Parallel interface: the data is spread over four parallel fiber links
- ▶ SW Serial interface: 10.2 Gb/s over a single fiber link
- ▶ SW Coarse Wavelength Division Multiplexed (CWDM): data is multiplexed over four wavelengths on a single fiber
- ▶ LW Serial interface: 10.2 Gb/s over a single fiber link
- ▶ LW CWDM: data is multiplexed over four wavelengths on a single fiber

The Fibre Channel Industry Association (FCIA) completed the core content of its proposed 10 Gb/s Fibre Channel standard. The forthcoming 10GFC specification leverages the work done by the IEEE P802.3ae Task Force and shares a common link architecture and common components with Ethernet and InfiniBand. The proposed 10GFC standard will span link distances from 15 m up to 10 km and offer direct support for native dark fiber (DWDM) and SONET/SDH, while preserving the Fibre Channel frame format and size for full backward compatibility.



SAN features

In this chapter we discuss some terminology and concepts that are derived from the Fibre Channel standards and frequently found in SAN device specifications and installations.

We also overview some common features and characteristics of the SAN environment, such as distance, applications, and the different platforms that can benefit from a SAN implementation.

3.1 Fabric implementation

We can build a SAN with a single switch and attached devices. However, as our fabric expands, we will eventually run out of ports. One possible solution is to move to a bigger switch or director, and another solution is to interconnect switches together to build a larger fabric. Another reason that we may need to interconnect switches or directors is to cover longer distances, for example, a building-to-building interconnection for backup and disaster recovery.

Note: It is not unusual to see directors referred to as switches. This is a statement as to the architecture that is employed *within* the director. That is to say, the director adheres to the Fibre Channel Switched Fabric (FC-SW) standard and employs the same switching protocol as a switch. There is no Fibre Channel “Director” Fabric standard! In this redbook, where something does not apply to both switches and directors equally, we make this distinction clear.

The diagram in Figure 3-1 shows two cascaded directors located at two different sites that can be up to 10 km apart. In this way all four servers can connect to both ESS devices.

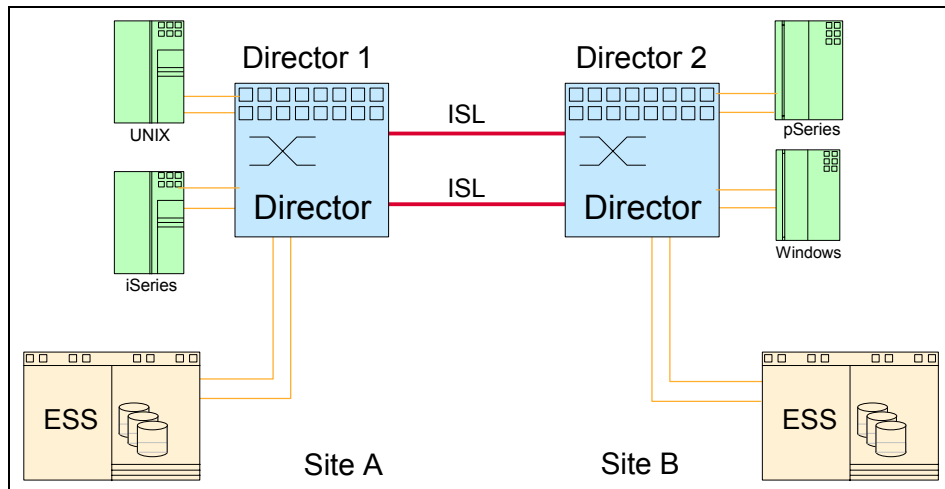


Figure 3-1 Cascading directors

3.1.1 Blocking

To support highly performing fabrics, the fabric components (switches or directors) must be able to move data around without any impact to other ports, targets, or initiators that are on the same fabric. If the internal structure of a switch or director cannot do so without impact, we end up with blocking.

Blocking means that the data does not get to the destination. This is opposed to congestion, where data will still be delivered, albeit with a delay. Switches and directors may employ a non-blocking switching architecture. Non-blocking switches and directors are the Ferraris on the SAN racetrack — they provide for multiple connections travelling through the internal components of the switch and director concurrently.

We illustrate this concept in Figure 3-2.

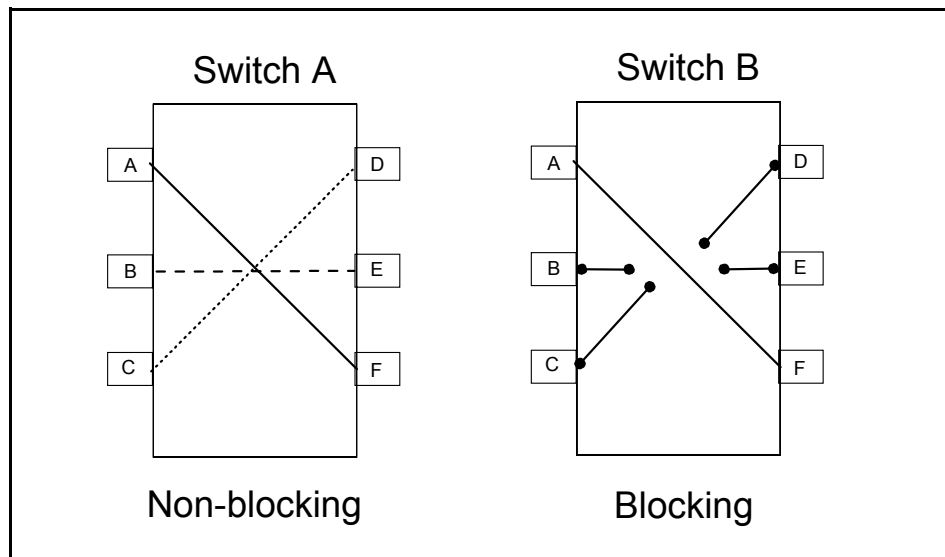


Figure 3-2 Non-blocking and blocking switching

In this example, non-blocking Switch A, port A speaks to port F, B speaks to E, and C speaks to D without any form of suspension of communication or delay; that is to say, the communication is not blocked. In the blocking Switch B, while port A is speaking to F, all other communication has been stopped or blocked.

3.1.2 Ports

The ports of a switch that connect to the devices N_Ports are called F_Ports. Coupling switches together introduces a new kind of connection, switch to switch. The port at which frames pass between switches within the fabric is called an E_Port.

A switch port will typically support one or more of the following Port Modes:

- ▶ F_Port (defined in the FC-PH standard)
- ▶ FL_Port (arbitrated loop connection defined in the FC-AL standard)
- ▶ E_Port (defined in the FC-SW standard).

A switch that only provides F_Ports and FL_Ports forms a non-expandable fabric. In order to be part of an expandable fabric the switch must have at least one port capable of E_Port operation.

A switch port that has the capability to support more than one port mode attempts to configure itself first as an FL_Port, then as an E_Port and finally as an F_Port, depending on which of the three modes are supported and the port it is connected to.

Switch ports that support both F_Port and E_Port modes are called G_Ports.

3.1.3 Inter-Switch Links

According to the FC-SW Fibre Channel standard, the link joining a pair of E_Ports is called an Inter-Switch Link (ISL).

ISLs carry frames originating from the node ports and those generated within the fabric. The frames generated within the fabric serve as control, management, and support for the fabric.

Before an ISL can carry frames originating from the node ports, the joining switches have to go through a synchronization process on which operating parameters are interchanged. If the operating parameters are not compatible, the switches may not join, and the ISL becomes “segmented”. Segmented ISLs cannot carry traffic originating on node ports, but they can still carry management and control frames.

Trunking

Depending on the estimated or measured traffic, you may connect some of your switches by parallel ISLs to share the load. The SAN standard routing protocol FSPF allows you to do so and use the cumulative bandwidth of all parallel ISLs (see Figure 3-3).

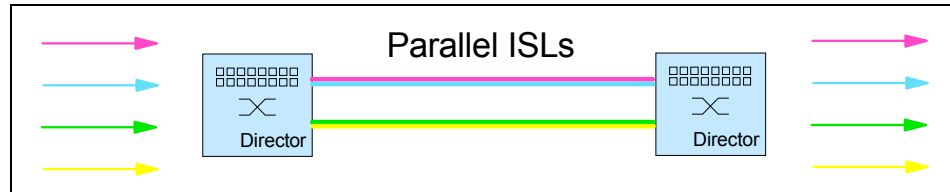


Figure 3-3 Parallel ISLs - low traffic

You need to be aware that load sharing reaches the boundary of its efficiency when servers send high amounts of data at the same time. As the switches dedicate the ISLs to the servers usually in a round-robin fashion, it may easily happen that one server occupies one ISL performing just a low rate of throughput and two other servers have to share the other ISL for their high rate of throughput (see Figure 3-4).

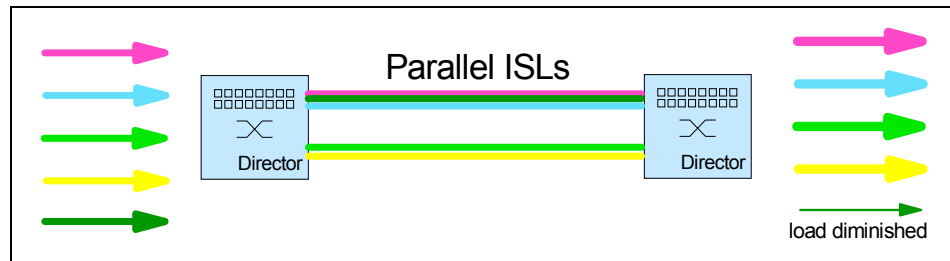


Figure 3-4 Parallel ISLs - high traffic

You may reduce, but not eliminate, this drawback by adding more ISLs in parallel; however, this may be far too expensive and subject to over-provisioning. Instead of this rather inflexible method of load *sharing*, switches may utilize a better way of load *balancing*. The implementation of load balancing is named trunking and is ideal for optimizing SAN performance (see Figure 3-5).

Each vendor of SAN switches will implement trunking in its own way. However, common to all their implementations is that transient workload peaks for one system or application are much less likely to impact the performance of other devices in the SAN fabric.

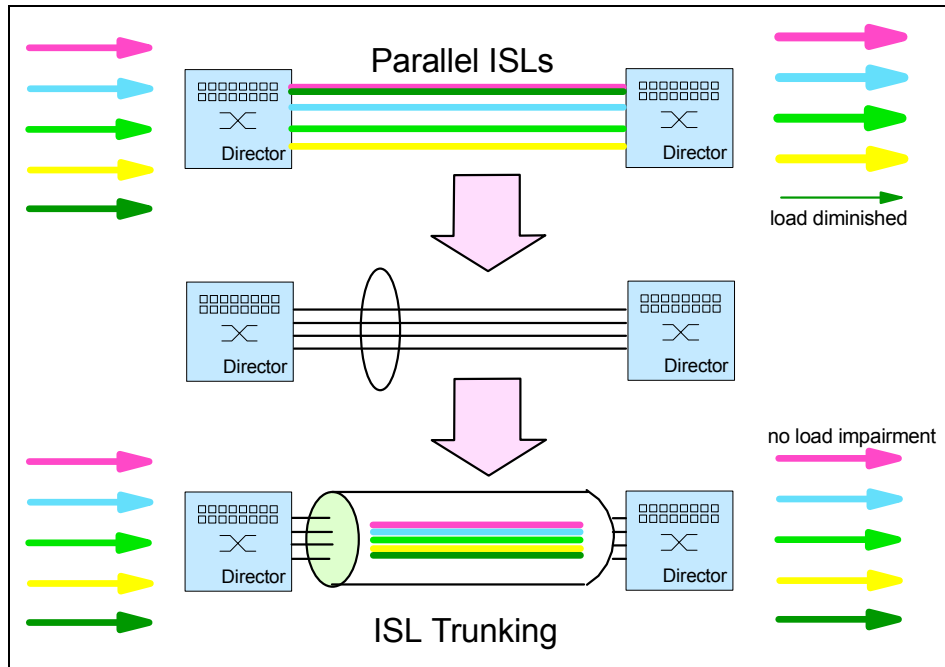


Figure 3-5 ISL Trunking

Load sharing or load balancing: Parallel ISLs always shared load or traffic in a “rough” server-oriented way: next server or next available ISL, regardless of the amount of traffic each server is causing. Load balancing provides the means to find an effective way to use all of the cumulative bandwidth of these parallel ISLs.

Oversubscribing the fabric

We can have several ports in a switch that can communicate with a single port, for example, several servers sharing a path to a storage device. In this case the storage path determines the maximum data rate that all servers can get, and this is usually given by the device and not the SAN itself.

When we start cascading switches, communication between switches are carried by ISLs, as previously stated. It is possible that several ports in one switch need to simultaneously communicate with ports in the other switch through a single ISL. In this case it is possible that the connected devices are able to sustain a data transfer rate higher than 100 MB/s, so the throughput will be limited to what the ISL can handle, and this may impose a throttle or roadblock within the fabric.

We use the term oversubscription to describe a situation when we have several ports trying to communicate with each other, and when the total throughput is higher than what that port can provide. Oversubscription, in itself, is not a bad thing. It is actually good, because it would be too cost prohibitive to dedicate bandwidth and resources for every connection. The problem arises if the oversubscription results in congestion. Congestion occurs when there is not enough bandwidth available for the application or connection. This can happen on storage ports and ISLs.

When designing a SAN, it is important to consider the possible traffic patterns to determine the possibility of oversubscription and which patterns may result in congestion. For example, traffic patterns during backup periods may introduce oversubscription that can affect performance on production systems. In some cases this is not a problem that may even be noticed at first, but as the SAN fabric grows, it is important not to ignore this possibility.

Fabric shortest path first

According to the FC-SW-2 standard, Fabric Shortest Path First (FSPF) is a link state path selection protocol. FSPF keeps track of the links on all switches in the fabric and associates a cost with each link. The protocol computes paths from a switch to all the other switches in the fabric by adding the cost of all links traversed by the path, and choosing the path that minimizes the cost.

For example, as shown in Figure 3-6, if we need to connect a port in switch A to a port in switch D, it will take the ISL from A to D. It will not go from A to B to D, nor from A to C to D.

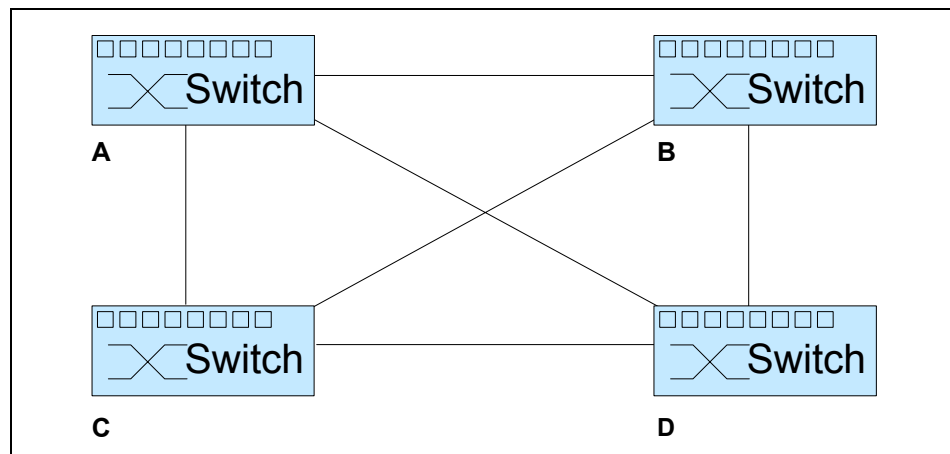


Figure 3-6 Four-switch fabric

This is because FSPF is currently based on the hop count cost.

The collection of link states (including cost) of all switches in a fabric constitutes the topology database (or link state database). The topology database is kept in all switches in the fabric, and they are maintained and synchronized to each other. There is an initial database synchronization, and an update mechanism. The initial database synchronization is used when a switch is initialized, or when an ISL comes up. The update mechanism is used when there is a link state change, for example, an ISL going down or coming up, and on a periodic basis. This ensures consistency among all switches in the fabric.

If we look again at the example in Figure 3-6, and we imagine that the link from A to D goes down, switch A will now have four routes to reach D:

- ▶ A-B-D
- ▶ A-C-D
- ▶ A-B-C-D
- ▶ A-C-B-D

A-B-D and A-C-D will be selected because they are the shortest paths based on the hop count cost. The update mechanism ensures that switches B and C will also have their databases updated with the new routing information.

Load balancing

The standard does not provide for load balancing when there are multiple paths of the same cost, so it is up to the switch vendor to establish routing algorithms to balance the load across ISLs. The potential routes are stored in routing tables.

Some vendors allow you to adjust the cost of traversing the fabric, and it is wise to check with each vendor as to the adjustments that can be made. Some vendors also allow you to define static routes. Again, it is wise to check with each vendor regarding what you can do to affect the traffic that flows over ISLs.

The balancing is usually done at initialization, assigning the same number of paths to each ISL. However, having the same number of paths does not mean having the same bandwidth requirements. We may end up with different connections that have high performance requirements being assigned to the same ISLs; while other ISLs are not being used due to inactive connections. Current implementations do not include dynamic load balancing, although this is expected to change over time.

Due to the potential performance impact of oversubscribing ISLs, it is recommended to have high volume traffic inside a switch or director. When cascading is not an option, the number of ISLs should be planned, and should take into consideration the expected traffic through them under different conditions, for example, production workload, and backup workload. In the absence of quantitative data, if you plan for the peak workload, that may be as good a rule of thumb as any.

When ISL oversubscription is detected, one solution is to add additional ISLs. It can be done concurrently, and the new path will be automatically included in the routing tables.

3.1.4 RSCN

The Registered State Change Notification (RSCN) is part of the Extended Link Service (ELS) in the Fibre Channel protocol. It was defined within the Fabric Loop Attachment group (FC-FLA) as a replacement for State Change Notification (SCN). You may consider RSCN similar to SCN plus the opportunity as a Fibre Channel device to register (subscribe) to that service or not. RSCN, like SCN, is used to notify FC devices about the status changes of other ports which may be of interest for them. For example, when a storage port becomes active or inactive, the switch will let the registered servers know by issuing a RSCN notification to them. RSCN notifications flow either from:

- ▶ **Node ports to switch:** By addressing the well-known fabric controller address of 0xFF FF FD (FC-FLA definition modified in FC-DA)
- ▶ **Switch to switch:** By addressing the fabric controller (FC-FLA definition modified in FC-MI-2)
- ▶ **Switch to node port:** From fabric controller to node fabric address (FC-FLA definition modified in FC-MI-2)

After a server has been notified via RSCN that another SCSI storage device has come online, the server may try and attach to that storage by performing a login to it. Or if the server was notified that some storage has gone offline, the server may like to verify the current status of that device. Without RSCN, in the latter case, the server probably wouldn't find out until it was sending SCSI-READs or WRITEs to that storage. These are the types of RSCNs:

- ▶ **Fabric Format:** Sent when a zone configuration is activated or deactivated or when an ISL in a fabric goes up or down.
- ▶ **Port Format:** Occurs when a device logs in or out of a fabric.
 - Sent to local devices on the same switch.
 - Sent to remaining switches in the fabric.
- ▶ **Area Format:** Occurs when an entire arbitrated loop goes up or down.
- ▶ **Domain Format:** Occurs when a switch is added or removed from a fabric.

3.2 Classes of service

In Fibre Channel, we have a combination of traditional I/O technologies with networking technologies.

We need to keep the functionality of traditional I/O technologies to preserve data sequencing and data integrity, and we need to add networking technologies that allow for a more efficient exploitation of available bandwidth.

Based on the methodology with which the communication circuit is allocated and retained, and in the level of delivery integrity required by an application, the Fibre Channel standards provide different classes of service:

3.2.1 Class 1

In a Class 1 service, a dedicated connection between source and destination is established through the fabric for the duration of the transmission. Each frame is acknowledged by the destination device back to the source device. This class of service ensures the frames are received by the destination device in the same order they are sent, and reserves full bandwidth for the connection between the two devices. It does not provide for a good utilization of the available bandwidth, since it is blocking another possible contender for the same device. Because of this blocking and the necessary dedicated connections, Class 1 is rarely used.

3.2.2 Class 2

In a Class 2 service there is no dedicated connection; each frame is sent separately using switched connections allowing several devices to communicate at the same time. For this reason Class 2 is also called “connectionless”. Although there is no dedicated connection, each frame is acknowledged from destination to source to confirm receipt. The use of delivery acknowledgments in Class 2 allows for quickly identifying communications problems at both the sending and receiving ports. Class 2 makes a better use of available bandwidth since it allows the fabric to multiplex several messages on a frame by frame basis. As frames travel through the fabric they can take different routes, so Class 2 does not guarantee in-order delivery. Class 2 relies on upper layer protocols to take care of frame sequence. The use of acknowledgments reduced available bandwidth which needs to be considered in large scale busy networks.

3.2.3 Class 3

Like Class 2, there is no dedicated connection in Class 3, the main difference is that received frames are not acknowledged. The flow control is based on BB Credit, but there is no individual acknowledgement of received frames. Class 3 is also called “datagram connectionless” service. It optimizes the use of fabric resources, but it is now up to the upper layer protocol to ensure all frames are received in the proper order, and to request to the source device the retransmission of any missing frame. Class 3 is the commonly used class of service in Fibre Channel networks.

Note: Classes 1, 2, and 3 are well defined and stable. They are defined in the FC-PH standard.

Cisco switches and directors support Class 2 and Class 3 service.

3.2.4 Class 4

Class 4 is a connection oriented service like Class 1, but the main difference is that it allocates only a fraction of the available bandwidth of a path through the fabric that connects two N_Ports. Virtual Circuits (VCs) are established between N_Ports with guaranteed Quality of Service (QoS) including bandwidth and latency. The Class 4 circuit between two N_Ports consists of two unidirectional VCs, not necessarily with the same QoS. An N_Port may have up to 254 Class 4 circuits with the same or different N_Port. Like Class 1, Class 4 guarantees in-order frame delivery and provides acknowledgment of delivered frames, but now the fabric is responsible for multiplexing frames of different VCs. Class 4 service is mainly intended for multimedia applications such as video and for applications that allocate an established bandwidth by department within the enterprise. Class 4 was added in the FC-PH-2 standard.

3.2.5 Class 5

Class 5 is called isochronous service and it is intended for applications that require immediate delivery of the data as it arrives, with no buffering. It is not clearly defined yet. It is not included in the FC-PH documents.

3.2.6 Class 6

Class 6 is a variant of Class 1 known as multicast class of service. It provides dedicated connections for a reliable multicast. An N_Port may request a Class 6 connection for one or more destinations. A multicast server in the fabric will establish the connections and get the acknowledgment from the destination ports, and send it back to the originator. Once a connection is established it should be retained and guaranteed by the fabric until the initiator ends the connection. Only the initiator can send data and the multicast server will transmit that data to all destinations. Class 6 was designed for applications like audio and video requiring multicast functionality. It appears in the FC-PH-3 standard.

3.2.7 Class F

Class F Service is defined in the FC-SW and FC-SW2 standard for use by switches communicating through ISLs. It is a connectionless service with notification of non-delivery between E_Ports, used for control, coordination and configuration of the fabric. Class F is similar to Class 2 since it is a connectionless service, the main difference is that Class 2 deals with N_Ports sending data frames, while Class F is used by E_Ports for control and management of the fabric.

3.2.8 Communication

FC-2 is the protocol level that defines protocol signaling rules and defines the organization or structure of Fibre Channel communications. This structure allows for efficient flow control and allows the network to quickly identify where a network error is occurring. The following describes the levels within this structure, they are listed according to size from the largest to the smallest.

- ▶ **Exchanges:** The highest level Fibre Channel mechanism used for communication. An exchange contains one or more non-concurrent sequences being exchanged between a pair of Fibre Channel ports.
- ▶ **Sequences:** A sequence is a collection of frames related to one message element or information unit.
- ▶ **Frames:** A Fibre Channel frame consists of maximum 2112 bytes of data. It is considered as a basic unit of data transmission. It consists of a start delimiter, destination and source address, protocol metadata, data payload, CRC (error check value) and an end delimiter.
- ▶ **Words:** An addressable unit of data in memory. The smallest Fibre Channel data element consisting of 40 serial bits representing either a flag (K28.5) plus 3 encoded data bytes (10 encoded bits each) or four 10-bit encoded data bytes. An *ordered set* is a 4-byte transmission word that has the special character, K28.5 as its first character and 3 bytes used to define the meaning or function of the ordered set. They either identify the start of frame, the end of frame, or occur between Fibre Channel frames.

3.2.9 Solutions

The main support for Fibre Channel development came from the workstation market. While in the mainframe platform the I/O channels have evolved allowing storage attachment and sharing through multiple high speed channels, and fiber optic cabling was introduced with ESCON®, workstations have been using SCSI as the common interface for storage attachment.

For storage interconnections, a SCSI interface had been traditionally used, but as data volumes and performance requirements increased, SCSI limitations started to surface: bulky cables, shared bus architecture that limits performance due to bus arbitration, limited distance of up to 25 m, and limited addressing of up to 15 targets. The continual growth of storage capacity requirements, data sharing needs and performance issues, made clear that it was necessary to overcome SCSI limitations. IBM introduced the IBM Serial Storage Architecture, and the IBM 7133 Disk Storage that solved many of the limitations of SCSI devices.

Another solution was the Fibre Channel interface. Mapping SCSI over the Fibre Channel Protocol was the solution that allowed access to multiple storage devices, extended distances, reduce cable bulk, and sharing of devices. Initially Fibre Channel arbitrated loop (FC-AL) was implemented to connect disk devices to hosts, and provided many benefits like smaller cables and connectors, faster data transfers and longer distances. Today, the arbitrated loop solution may still work for a department or workgroup, but does not offer the performance and connectivity required by an enterprise SAN, so different vendors offer Fibre Channel HBAs which provide for point-to-point connection, as well as connections to a Fibre Channel switched fabric.

These are some of the many reasons to implement a SAN fabric implementation:

- ▶ **Storage consolidation:** Storage devices can be shared with more servers without increasing the number of ports in the device.
- ▶ **Clustering:** For high availability solutions, a SAN allows shared storage connections and provides for longer distances between devices.
- ▶ **LAN free backup:** The ability to consolidate tape drives and tape libraries and share them among several backup hosts provides the opportunity to optimize the utilization of the tape drives. The result is more data can be backed up with the same number of, or less, drives.

To expand the benefits of a SAN across longer distances and allow more companies to realize the benefits of a SAN, several projects are currently underway. One solution that may make use of Fibre Channel is the Internet Protocol (IP). It requires an upper layer protocol that takes care of sending IP packets as Fibre Channel Sequences. One project in the T11 committee deals with Fibre Channel link encapsulation (FC-LE). As a result of this project's work, there is a new protocol known as FC-IP, or FCIP, that will allow for greater distances by IP encapsulating the Fibre Channel protocol.

A new protocol called iSCSI will enable SCSI commands to be packaged and sent over existing IP networks. This will allow companies that currently do not have the resources for a Fibre Channel network to build a SAN utilizing their existing IP network.

Example applications that exploit SANs

Today, we have information available in many different forms: text, images, audio, and video, which we usually refer to as multimedia. Given the storage capacity and performance levels that computer systems are able to provide today, and what can be expected in the future, it is becoming practical to store, distribute, and retrieve more information in digital form. This can dramatically increase the amount of data stored, the transmission throughput, and the sharing requirements.

SANs can fulfill many of the demands of multimedia applications. Some examples of these are described in the following sections.

Video editing

Digital video editing is usually performed in standard computers with specialized video boards. Due to the size of the video files they have to handle, workstations normally have big and fast local storage devices but are interconnected by a relatively low bandwidth LAN. Local copies must be made of the files to be worked on, and once again when the work is finished, either from a server through the LAN or using removable media.

Interconnecting all workstations and storage by means of a SAN has the following advantages:

- ▶ **Bandwidth:** A properly designed SAN can provide enough bandwidth for the workstations to access the data in shared storage eliminating the need and the time required for local copies.
- ▶ **Storage efficiency:** Consolidating storage in a single pull allows each workstation have the amount required for each task eliminating excess storage in each workstation.
- ▶ **Reliability:** Installing fault tolerant storage like RAID arrays and being able to perform simultaneous centralized backups and restores.
- ▶ **Workstation flexibility:** A task initiated in one workstation can be continued in another one. It is also possible to introduce a new platform and share existing data. This also eliminates the workstation as a single point of failure.
- ▶ **Location flexibility:** SAN distances allow workstations to be far apart from each other, making it possible to have some workstations in special locations, like soundproof rooms.
- ▶ **Simultaneous viewing:** The work in progress can be monitored without interruptions to the creative staff.
- ▶ **Separation of duties:** Creative staff can concentrate in their work leaving repetitive tasks like digitizing material and recording output to tape to other personnel. This can also help consolidating and saving additional equipment like VCRs.

In order to allow several users to share the storage and working copies simultaneously, some kind of management software is required.

Pre-press

Some of the characteristics of today's graphic industry are:

- ▶ **Multiple parties involved:** Printing facilities, advertising agencies, graphic designers, clients.
- ▶ **Gigantic size files:** Raster image processors (RIPs) and open pre-press interface (OPI) applications that feed computer to plate (CTP) devices often generate files in the gigabyte range, especially when dealing with large images at 1200 dpi resolution, making LANs too slow for reasonable file transfer times.
- ▶ **Mixed platforms:** High end graphic workstations running on Apple Macintosh workstations running Windows NT® or 2000, application and file servers running UNIX.

LANs have been successfully handling the mixed platform environment, but as file sizes have increased, LAN speed has become the bottleneck.

SANs can also handle the mixed platform environment and not only offer a greater bandwidth, but also reduce the processor overhead to move the larger data files. This makes it possible to connect the different workstations directly to shared storage and eliminates the need for making local copies of large files.

Video distribution

There is a lot of research in process related to multimedia systems. It deals with storing and transmitting large amounts of time critical data between storage systems and end users.

Some of the characteristics of a video server are:

- ▶ Huge amount of data stored
- ▶ Large number of potential concurrent users
- ▶ Real time requirements to allow all users receive a jitter free video signal
- ▶ User interaction, such as title selection, play, pause, stop
- ▶ Availability 24 hours x 365 days required
- ▶ Multiple server configurations to be able to handle all potential users
- ▶ Multiple storage devices to handle the amount of data stored

Without entering into the design considerations of such a system, SANs offer the capacity of handling the multiple server, multiple storage configuration with the bandwidth, the high availability, and the scalability characteristics required by these kinds of applications.

3.3 Distance

Fibre Channel allows for much longer distances than the 25 m limit of SCSI links. Currently supported distances are:

- ▶ **Short-wave laser:** Up to 500 m (50 μ m fiber), 300 m (62.5 μ m fiber)
- ▶ **Long-wave laser:** Up to 10 km (9 μ m single-mode fiber).

When longer distances are required, there are different options, for example, extenders, protocol converters, or Dense Wave® Division Multiplexors (DWDM), and selection will depend on the available links between the two locations, distance and budget.

The chart in Figure 3-7 compares the distances that can be reached by different alternatives.

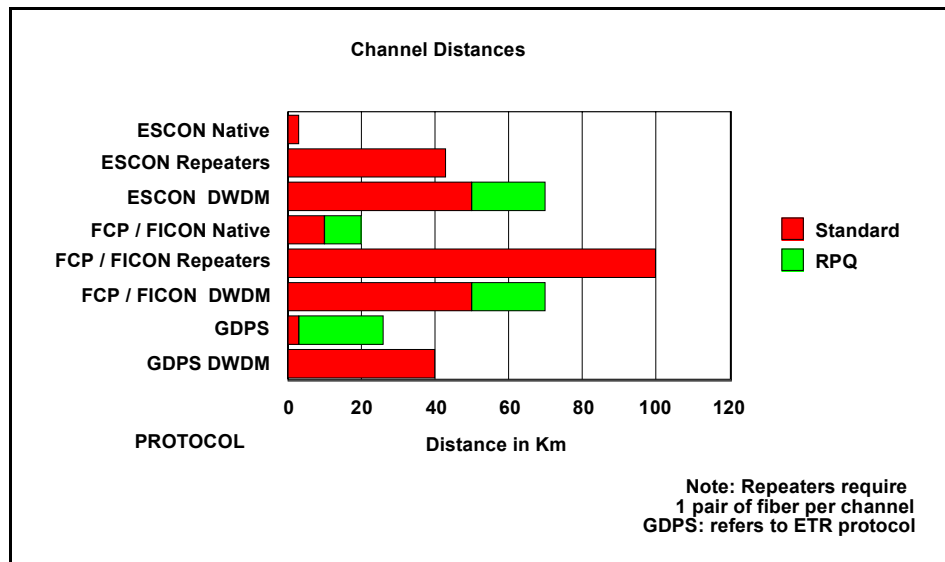


Figure 3-7 Channel distances comparison

Some distance solutions convert FCP to several OC3 or ATM channels, route the signals through telco lines, and reconvert the signals at the other end. These can reach hundreds or thousand of miles. By using repeaters and dedicated fibers, we can get distances of about 100 km. DWDM allows us to send several channels over the same fiber.

Figure 3-8 shows an example of a distance solution for tape backup. SCSI and FC links are multiplexed and transmitted over Telco lines to the remote site.

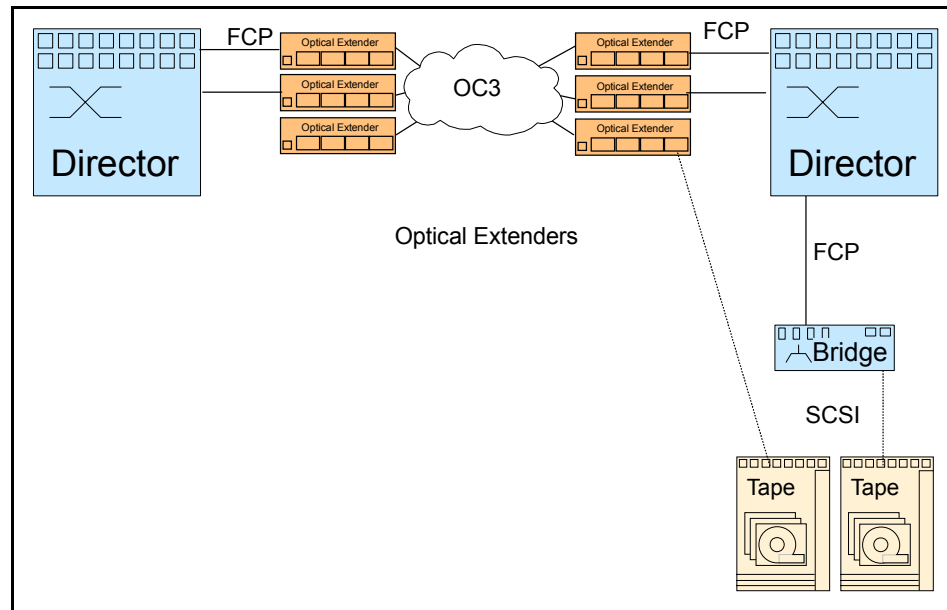


Figure 3-8 Distance solution

3.3.1 Dark fiber

Dark fiber is a dedicated, end-to-end, fiber optic cable that can be used without additional equipment up to 10 kilometers for longwave transceivers, or may require the use of extenders or repeaters, either external or internal in some directors, for longer distances. By using dark fiber, we can get the most direct connection and full bandwidth, but the down side is the cost of the dedicated fiber links.

Figure 3-9 shows an example of a solution using Finisar extenders. It is purely a fiber solution, but requires a pair of dark fibers between the sites for each link.

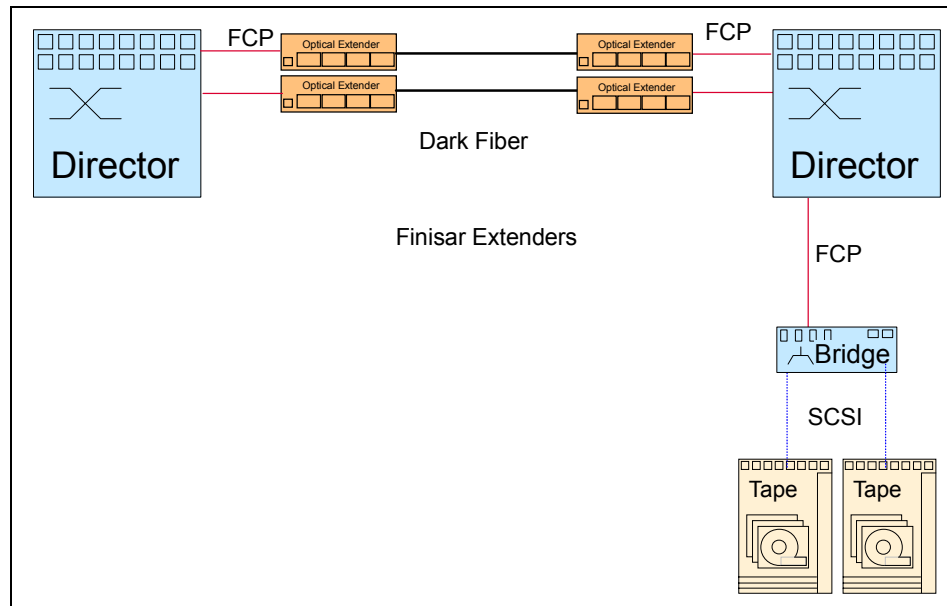


Figure 3-9 Distance solution with Finisar extenders

3.3.2 Dense Wavelength Division Multiplexing

Dense Wavelength Division Multiplexing (DWDM) allows several fiber optic signals to be multiplexed and sent over the same fiber optic cable at long distances reducing cabling requirements.

The original fiber optic signals are converted into electrical signals. The electrical signals are then converted back to different wavelength fiber optic signals. At the receiving end, the signals are optically filtered, converted back to the original signal type, and sent to the connecting device.

In Figure 3-10 we show an example of DWDM point-to-point configuration.

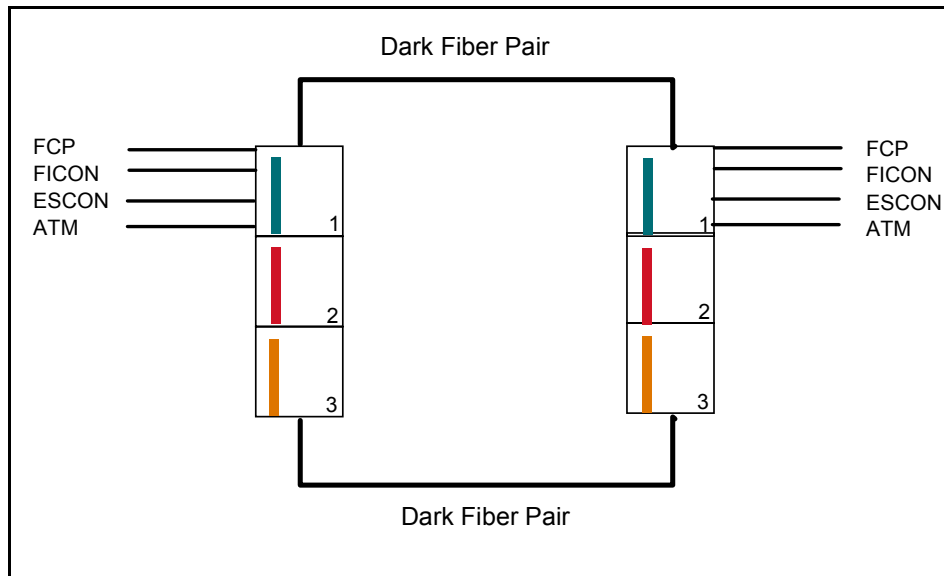


Figure 3-10 DWDM point-to-point configuration

To ensure redundancy and high availability, the two cables must go through completely different paths. They must enter the building at different entry points and follow different channels inside the building. When going out of their own building and campus, consider obtaining the links through different carriers.

In contrast to Figure 3-10 on page 103, which shows a simplified representation of a point-to-point configuration; the other possible configuration is the hubbed ring configuration. This is shown in Figure 3-11.

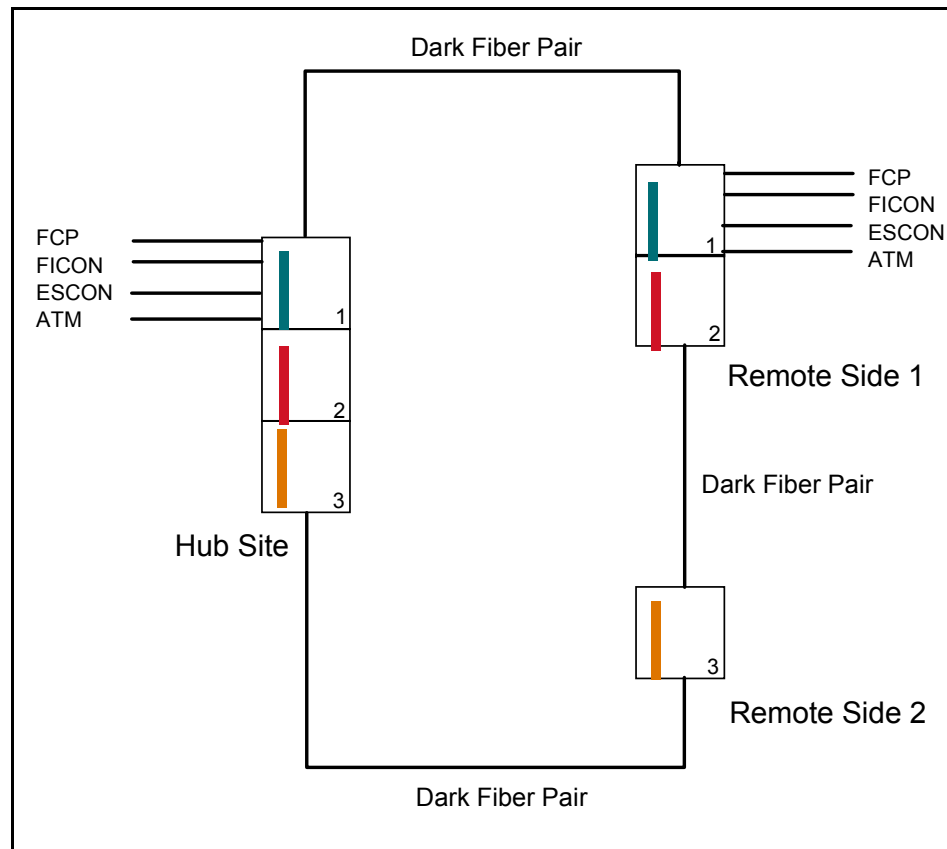


Figure 3-11 DWDM hubbed ring configuration

3.3.3 Primary and secondary routes

From an availability standpoint, it is essential that redundant links do not have common points of failure. It is very important that the carrier knows the physical routes the links are following, and that they are not only independent with regard to the transmission equipment, but also geographically apart. This introduces the possibility of different link lengths and different signal travel times.

We discuss the concepts associated with SAN distance solutions in *Introduction to SAN Distance Solutions*, SG24-6408.

3.4 Time-out values

Longer distances introduce other factors to consider in the SAN design, one of which is latency. The latency increases due to the time needed for the signal to travel the longer links, and has to be added to the normal latency introduced by switches and/or directors. Another point is that the time-out values should allow for increased travel times. For this reason, parameters such as the E_D_TOV and R_A_TOV have to be evaluated.

The FC-PH standard defines three time-out values used for error detection and recovery:

R_T_TOV

This is the Receiver Transmitter time-out value. It is used by the receiver logic to detect Loss of Synchronization with the transmitter. It has a fixed value of 100 ms.

E_D_TOV

This is the Error Detect time-out value. It represents the period in which a response should come back for a timed event. For example, during data transmission it represents a time-out value for a data frame to be delivered, the receiving port to transmit a response and the response be received by the initiator. E_D_TOV can normally be configured. The selected value should consider configuration and switch characteristics.

E_D_TOV is used in class of services 1 and 2, since class 3 does not check for acknowledgment.

R_A_TOV

This is the Resource Allocation time-out value. It is used as a time-out value during the recovery process. It should be set to E_D_TOV plus twice the maximum time a frame may be delayed within a fabric and still be delivered.

3.4.1 Time-out value settings

Without entering into the details of error detection and recovery, it is important to know the consequences of a wrong time-out value setting. Small E_D_TOV values may affect performance due to sequences being timed out and retried when they can still be correctly finished; too small R_A_TOV values may cause duplicated frames during recovery. On the other hand, if the values are too long, error detection and recovery may be delayed when it is needed.

Switch manufacturers provide default values that should work fine for normal distances (up to 10 km). Delay considerations should be taken into account for extended distances. Each kilometer of fiber adds approximately 5 microseconds delay. Also the delay introduced by repeaters or extenders should be considered.

It is wise to check with each vendor as to what these values should, or need, to be set to in a fabric.

3.5 Buffers

Ports need memory, or “buffers”, to temporarily store frames as they arrive and until they are assembled in sequence, and delivered to the upper layer protocol.

The number of buffers (the number of frames a port can store) is called its “Buffer Credit”.

BB_Credit

During login, N_Ports and F_Ports at both ends of a link establish its Buffer to Buffer Credit (BB_Credit).

EE_Credit

During login all N_Ports establish End to End Credit (EE_Credit) with each other.

During data transmission, a port should not send more frames than the buffer of the receiving port can handle before getting an indication from the receiving port that it has processed a previously sent frame. Two counters are used for that purpose. BB_Credit_CNT and EE_Credit_CNT, and both are initialized to 0 during login.

Each time a port sends a frame, it increments BB_Credit_CNT and EE_Credit_CNT by 1. When it receives R_RDY from the *adjacent* port it decrements BB_Credit_CNT by 1, when it receives ACK from the *destination* port it decrements EE_Credit_CNT by 1. Should at any time BB_Credit_CNT become equal to the BB_Credit, or EE_Credit_CNT become equal to the EE_Credit of the receiving port, the transmitting port has to stop sending frames until the respective count is decremented.

The previous statements are true for Class 2 service. Class 1 is a dedicated connection, so it does not care about BB_Credit and only EE_Credit is used (EE Flow Control). Class 3 on the other hand is an unacknowledged service, so it only uses BB_Credit (BB Flow Control), but the mechanism is the same on all cases.

Here we can see the importance that the number of buffers has in overall performance. We need enough buffers to make sure the transmitting port can continue sending frames without stopping in order to use the full bandwidth. This is particularly true with distance.

BB_Credit considerations for long distance

BB_Credit needs to be taken into consideration on Fibre Channel devices that are several kilometers apart from each other and you need to know the distance separating the adjacent partners. We will assume that Fibre Channel devices *A* and *B* are connected by a 10 km fiber optic cable, as shown in Figure 3-12.

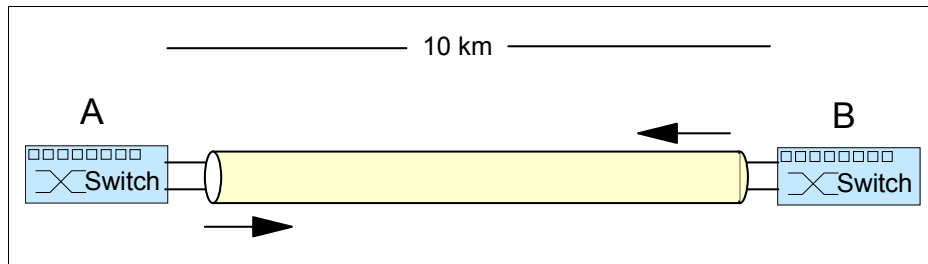


Figure 3-12 Adjacent FC devices

Light travels at approximately 300,000 km/s through a vacuum and at about 200,000 km/s through glass fiber. Latency is the inverse function of speed — so the optical signal of a Fibre Channel frame ends up with a latency of 5 ns/m.

$$latency = \frac{1}{speed} = \frac{1}{0.2 \exp 9 \frac{m}{s}} = 5 \exp -9 \frac{s}{m} = 5 \frac{ns}{m}$$

Light has a finite speed, and we need to take that into account when we figure out the maximum amount of frames that will be in transit over the fiber optic cable from *A* to *B*. A distance of 10 km over Fibre Channel means a round-trip of 20 km in total. The optical signal takes 100 μs propagation time (t_p) to travel that round-trip.

$$t_p = distance \times latency = 20 \exp 3 m \times 5 \exp -9 \frac{s}{m} = 100 \exp -6 s = 100 \mu s$$

That is the shortest possible time that *A* can expect to get an R_RDY back from *B* which, once received, signals that more frames can be sent.

Round-trip: We assume a data frame is sent from *A* and when it arrives at *B* a Receiver Ready (R_RDY) travels back to *A*. So for our equations, it is based on one frame which would make the round-trip.

Fibre Channel frames are usually 2 KB large, but because of 8b/10b encoding, they will become larger, as the encoding causes 1 byte to become 10 bits. Sending 2 KB Fibre Channel frames over fiber optic cable with 1 Gb/s bandwidth computes to a sending time (t_s) of 20 μ s per frame. In other words: 20 μ s is the time *A* needs to send 2000 bits.

$$t_s = \frac{\text{Framelength}}{\text{Bandwidth}} = \frac{20 \exp 3 \text{ b}}{1 \exp 9 \frac{\text{b}}{\text{s}}} = 2 \exp -6 \text{ s} = 20 \mu \text{ s}$$

To give an idea of how long a Fibre Channel frame spreads out on the fiber optic cable link, in 20 μ s the light travels 4000 m (l), so the 2 KB frame occupies 4 km of fiber optic cable from the first bit transmitted to the last bit transmitted.

$$l = t_s \times \text{speed} = 20 \exp -6 \text{ s} \times 0.2 \exp 9 \frac{\text{m}}{\text{s}} = 4 \exp 3 \text{ m} = 4000 \text{ m}$$

The ratio between propagation time and sending time gives us the maximum amount of frames which *A* may send out before *B*'s response would arrive back at *A*.

$$BB_Credit = \frac{t_p}{t_s} = \frac{100 \mu \text{ s}}{20 \mu \text{ s}} = 5$$

So, *A* may send out five consecutive frames to fill up the whole Fibre Channel during the time it is waiting for a response from *B*. In order to do so, *A* needs to hold at least 5 BB_Credits to use the Fibre Channel effectively. Distances in the range of a few hundred meters and below are not usually effected. It becomes an area for consideration with longer distances in the region of 50-100 km or more when extending Fibre Channel links over DWDM or ATM. To guarantee the same effectiveness for the optical signal's propagation time for a 100 km distance, you would need to make sure that *A* has 50 BB_Credits available. Doubling the bandwidth of the fiber optic link from 1 Gb/s to 2 Gb/s means there may be twice as many Fibre Channel frames on the link at the same time and so we will need twice as many BB_Credits to reach the same efficiency. That is theoretically 100 BB_Credits on a 100 km, 2 Gb/s fiber optic link.

Practically speaking, you may not need that much BB_Credit, as it is unlikely that the FC device will fill up the Fibre Channel 100% over a sustained period.

3.6 Data protection

Since data is the most valuable asset for any organization and data availability is a primary concern, independently of the measures taken to ensure backup and recovery it is fundamental to maximize data availability for, and from, the primary storage device.

3.6.1 RAID

One of the common methods used to protect data in case of disk drive failures is RAID (Redundant Array of Independent Disks).

RAID is an architecture designed to improve data availability by using arrays of disks together with data striping methodologies.

There are different RAID levels according to the methodology implemented. The original classification described RAID-1 to RAID-5. Later on, other levels were added like RAID-0, RAID-0+1, RAID-6 and RAID-10. Listed below are the most prevalent RAID types in use today:

- ▶ **RAID 0 (striping):** Striping of data across separate disks. RAID-0 does not provide any redundancy, it is only used for performance.
- ▶ **RAID 1 (mirroring):** Two copies of the data are written to separate disks. In case one disk fails, the data is still available from the other copy. Twice the number of disk drives are required to store the data.
- ▶ **RAID 5:** An array of n drives is formed. Records of data are striped in $n-1$ drives, parity generated and written on the remaining drive. There is no dedicated parity drive, data and parity are interleaved in all disks. In case of a disk failure, data in the failed disk is reconstructed from the remaining disks reversing the parity algorithm. It requires less disk drives than RAID-1, but there is a write penalty associated with generating parity for each record, since old data and parity must be read, new parity generated, and new data and parity written for each update. On the other hand, it allows better read performance enabling access to several disks at the same time. It is best for systems in which performance is not critical or which do few write operations.

RAID types such as 0+1 or 1+0 refer to the RAID implementation that is done first. RAID 1+0 would offer more protection than RAID 0+1.

Table 3-1 provides a summary of RAID level definitions.

Table 3-1 RAID levels definitions

RAID level	Description
RAID 0	Data striping, no fault protection.
RAID 1	Disk mirroring. Dual copy
RAID 0+1	Data striping and mirroring
RAID 2	Bit interleave with hamming code
RAID 3	Bit interleave data striping with parity disk
RAID 4	Block interleave data striping with parity disk
RAID 5	Block interleave data striping with skewed parity disk
RAID 6	Block interleave data striping with double skewed parity disk
RAID1+0(RAID10)	Data mirroring and striping

Some disk storage subsystems offer RAID protection without the need for operating system intervention. For example, the IBM Enterprise Storage Server (ESS) can be configured for RAID 5 or RAID 10 without any performance degradation, since it is controlled by the internal software.

3.6.2 Mirroring

The basic idea of mirroring is to preserve data availability by having two or more copies in different storage devices. If one device fails, the data is still available from another device.

Mirroring can be performed by hardware, like disk arrays supporting RAID-1 or ESS Peer-to-Peer Remote Copy (PPRC); by software, like AIX Logical Volume Manager Mirroring, Veritas Volume Manager, and Windows NT mirroring; or by a combination of hardware and software, like ESS Extended Remote Copy (XRC) in the OS/390® platform.

We can also differentiate between synchronous or asynchronous mirroring. In synchronous mirroring, any write must be completed on all copies before the operation is considered complete. In asynchronous mirroring, the write operation is completed in the primary device and then propagated to the copies.

Synchronous mirroring allows a quick data recovery, since both copies are exactly the same, but it has an impact in write performance, since we have to wait until write operations are performed in all devices.

Asynchronous mirroring does not have the write penalty, but in the case of a failure, we may lose the updates that are pending, so it has an impact on the time and procedures required for data recovery when needed.

The distance supported by SAN allows us to have devices further apart from each other and that way we can mirror a storage device in one site with another located in a different site. Having storage devices mirrored at different locations preserves data availability — not only in the case of disk failures, but also in case of any disaster affecting the primary site.

3.6.3 Clustering

Basically, a cluster is a group of servers that appear to clients in the network as a single entity. This group is managed as a single machine and the physical characteristics are transparent to users. Each individual server is known as a node.

The main benefits of clustering are scalability and availability. When workload increases, nodes can be added to absorb the additional workload and keep the performance levels. In the same way, if a node fails, or it is too busy, its workload can be absorbed by the remaining nodes transparently to the end users.

There are different classifications of clustering according to the way it is implemented: there are clusters implemented at a hardware level, at the operating system level, and the application level.

What all clustering solutions have in common is the requirement for shared access to storage. Here is where a SAN comes in to play, by offering the capacity of attaching different servers and different storage devices, providing the required bandwidth for concurrent access to data, and supporting the scalability requirements of adding additional servers or storage devices when required without modifying the basic infrastructure.

The Fibre Channel technology also offers the possibility of a longer distance between nodes or between nodes and storage devices.

A geographically dispersed cluster can provide a business continuity solution in case of a disaster affecting one of the sites.

3.6.4 Dual pathing

The idea of dual pathing, or multipathing in general, is to provide for a higher bandwidth so more data transfers can take place simultaneously, and also to maintain data availability in case of path failures.

Multipathing is common in the OS/390 environment, and the channel subsystem takes care of it. In the open systems environment, we get an instance of each device on each path, so the appropriate software is required to handle multipath configurations.

Since a SAN typically provides more than one path between a server and a storage device, multipathing software is mandatory.

Different vendors provide their own version of multipath software. Both Emulex and QLogic also offer multipathing with the purchase of their HBAs.

IBM has been offering the Data Path Optimizer (DPO) for AIX and Windows NT. For all ESS customers DPO has been superseded by the IBM Subsystem Device Driver (SDD).

IBM Subsystem Device Driver

The IBM Subsystem Device Driver (SDD) resides in the host server with the native disk device driver for the ESS. It uses redundant connections between the host server and disk storage in an ESS to provide enhanced performance and data availability.

The IBM Subsystem Device Driver provides the following functions:

- ▶ Enhanced data availability
- ▶ Automatic path failover and recovery to an alternate path
- ▶ Dynamic load balancing of multiple paths
- ▶ Concurrent download of licensed internal code
- ▶ Path selection policies for the AIX operating system:
 - IBM AIX
 - HP
 - Linux Red Hat and SuSE (Intel®)
 - Novell
 - Sun
 - Microsoft® Windows NT and 2000

In most cases, host servers are configured with multiple host adapters with SCSI or Fibre Channel connections to an ESS that, in turn, provides internal component redundancy. With dual clusters and multiple host interface adapters, the ESS provides more flexibility in the number of I/O paths that are available.

When there is a failure, the IBM Subsystem Device Driver reroutes I/O operations from the failed path to the remaining paths. This function eliminates the following connections as single points of failure: a bus adapter on the host server, an external SCSI cable, a fiber-connection cable, or a host interface adapter on the ESS. This automatic switching in case of failures is called path failover.

In addition, multi-path load balancing of data flow prevents a single path from becoming overloaded with I/O operations.

For the specific versions that are supported, and additional information about SDD refer to the *IBM Subsystem Device Driver Users Guide*, SC26-7478 at Web site:

<http://ssddom02.storage.ibm.com/techsup/webnav.nsf/support/sdd>

3.7 SAN platforms

In the topics that follow we discuss some of the platforms that are encountered in a SAN:

- ▶ zSeries® and S/390
- ▶ pSeries®
- ▶ xSeries®
- ▶ iSeries™

3.7.1 zSeries and S/390

The zSeries and S/390 platforms have a dedicated I/O subsystem that offloads workload from the processors allowing for high I/O data rates. Installations with high numbers of I/O devices and channels are very common. To solve the cabling and distance limitations of the Bus and Tag I/O connections, ESCON was introduced more than 10 years ago. ESCON, a serial interface using fiber optics as connecting media, has been delivering customers increased distance, reduced cable bulk, disk and tape pooling, clustering and data sharing, while providing management capabilities. Other vendors also supported ESCON, and it was adopted as a standard by NCITS.

FICON™ comes as an ESCON evolution. FICON is based on the Fibre Channel standard, so OS/390 and z/OS® is positioned to participate in heterogeneous Fibre Channel based SANs.

FICON support started bridging from FICON channels to existing ESCON directors and ESCON control units, delivering value using channel consolidation, cable reduction, increased distance and increased device addressability.

We are now in the next phase that includes native FICON control units, attached either point-to-point, or switched point-to-point, using FICON capable directors. Since FICON is an upper layer protocol using standard Fibre Channel transport, FICON directors are highly available Fibre Channel switches with capabilities that allow in-band management.

From an availability point of view, zSeries and S/390 offer the possibility of a Parallel Sysplex® configuration, the highest available configuration in the market. In a Parallel Sysplex, several processors are normally sharing I/O devices. ESCON directors have traditionally been used to allow sharing while reducing the number of control unit ports and the cabling requirements. With the introduction of FICON, processors and I/O can share a SAN with other platforms.

More information about Parallel Sysplex can be found at this Web site:

<http://www-1.ibm.com/servers/eserver/zseries/psa/>

Additional information about SAN on the zSeries and S/390 platforms, and future directions can be found at this Web site:

<http://www-1.ibm.com/servers/eserver/zseries/san/>

3.7.2 pSeries

There are different vendors that offer their own version of UNIX in the market. Each vendor (IBM, SUN, HP) offers its own hardware and different flavors of the UNIX operating system (AIX, Solaris, HP UX), each having some unique enhancements and often supporting different file systems (JFS, AFS®).

There are also several versions of management software available for the UNIX environment, such as Tivoli Storage Manager (TSM), and Veritas.

The IBM version of UNIX is the AIX Operating System and pSeries hardware. IBM currently offers a wide range of SAN ready pSeries servers from the entry servers up to Large Scale SP systems. Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redbook, *Practical Guide for SAN with pSeries*, SG24-6050.

More details can be found at the Web site:

<http://www-1.ibm.com/servers/solutions/pseries/>

3.7.3 xSeries

The platform of Intel based servers running Windows is a fast growing sector of the market. More and more of these servers will host mission critical applications that will benefit from SAN solutions such as disk and tape pooling, tape sharing, and remote copy.

IBM offerings in this platform include Netfinity®, NUMA-Q®, and xSeries. Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redpaper extract, *Implementing IBM server xSeries SANs*, REDP0416.

More details regarding xSeries can be found at the Web site:

<http://www-1.ibm.com/servers/solutions/xseries/>

3.7.4 iSeries

The iSeries platform uses the concept of single-level storage. The iSeries storage architecture (inherited from its predecessor systems System/38™ and AS/400®) is defined by a high-level machine interface. This interface is referred to as Technology Independent Machine Interface (TIMI). It isolates applications and much of the operating system from the actual underlying systems hardware. They are also unaware of the characteristics of any storage devices on the system because of single-level storage.

The iSeries is a multi-user system. As the number of users increase, you do not need to increase the storage. Users share applications and databases on the iSeries. As far as applications on the iSeries are concerned, there is no such thing as a disk unit. The idea of applications not being aware of the underlying disk structure is similar to the SAN concept.

Additional information regarding connecting pSeries servers to a SAN can be found in the IBM Redbook, *iSeries in Storage Area Networks: Implementing Fibre Channel Disk and Tape with iSeries*, SG24-6220-00.

More details regarding iSeries can be found at the Web site:

<http://www-1.ibm.com/servers/solutions/iseries/>

3.8 Security

Today, Fibre Channel fabrics are deployed in a variety of applications. Being developed to obtain SCSI connectivity for storage applications, Fibre Channel removes the traditional boundaries associated between SCSI devices. Storage interfaces, once propagated within the server or data center, may be extended outside the typically secure physical cabinet or location.

The Fibre Channel fabric and its components are also considered shared resources. This includes the switched fabric, the storage attached to the fabric and the services the fabric provides. This accessibility, plus the increase in both the quantity of data and users of that data, has heightened the awareness of the IT community to security exposures. As such, Fibre Channel fabrics and their components are prone to the same types of security breaches once associated mostly with IP networks.

3.8.1 Control types

Many of the security layers can be configured to ensure that only the appropriate users or systems have access to data. Determining the specific values for these different variables and then correctly implementing steps to enforce the policies can be quite challenging. The following three control types are important to creating a complimentary matrix of checks and balances to gain security in a SAN:

- ▶ **Operational controls:** Processes that are typically implemented by security personnel. These controls have to work well in conjunction with management controls and technical controls, and are defined in security documents. refer to “Operational controls” on page 143 for more information on this type.
- ▶ **Management controls:** Procedures and policies that ensure that there is proper management oversight with regards to security.
- ▶ **Technical controls:** Security measures implemented by hardware and software. In conjunction with strong operational controls and management controls, technical controls can detect unauthorized access, track changes, implement security policies.

More information about policies can be found at the Web sites:

<http://www.sans.org/resources/policies/>

<http://www.ait.s.uillinois.edu/security/securestandards.html#introduction>

3.8.2 Vulnerabilities

It is a matter for debate if there is such a thing as 100% security for any kind of data. You could disconnect the information database from the rest of the world and “seal” it in concrete to preserve it — but then it would be rather inaccessible and probably not of much value in such an environment. Accessibility for your business applications on SAN will be achieved by interconnection and the network. The network, being a LAN or a SAN with hundreds or more servers, storage devices, and connections, can have plenty of weaknesses in security to consider. You may reduce the possible vulnerabilities into three groups categorized by traffic type, as shown in Figure 3-13.

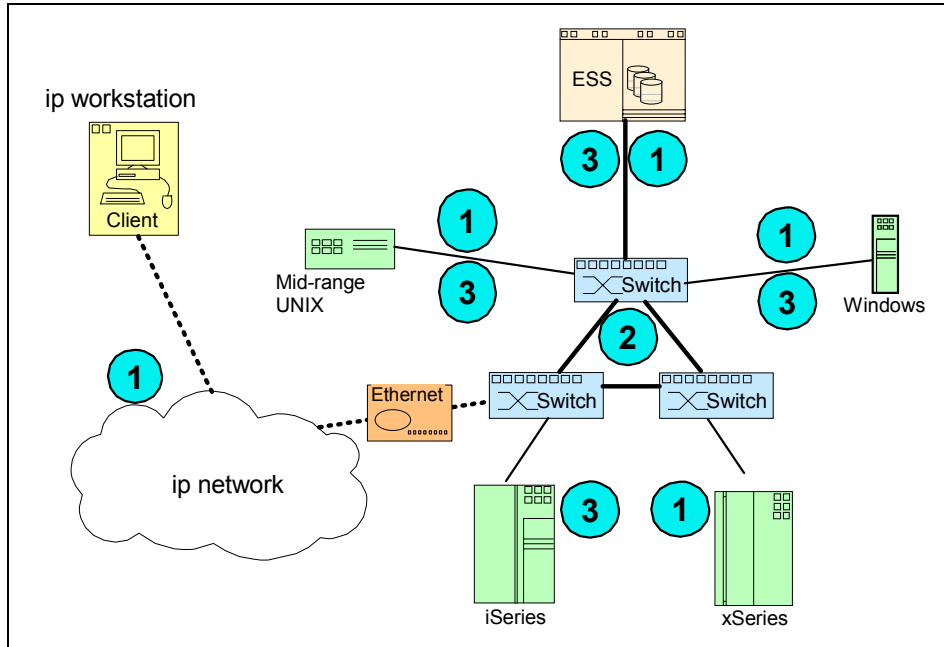


Figure 3-13 Vulnerabilities

Device-fabric traffic

This category includes any traffic that originates on servers and terminates on a fabric switch using Fibre Channel, IP, or other means. The devices can be systems that plug into the Fibre Channel ports of the SAN fabric via an HBA, as well as workstations connected through LANs and used by SAN administrators to manage the SAN. This type of risk can include traffic from unauthorized servers connected to the wrong port, invalid management device connections, and attacks such as denial of service (DOS).

Inter-switch traffic

Inter-switch traffic risks are limited to issues posed by the interconnection of SAN switches in a fabric. Inter-switch traffic includes the traffic generated when switches attempt to create E_Ports, as well as any traffic originating *and* terminating on a fabric switch. This category does not include traffic that only traverses the switches, such as data from a server to a storage device. The concern here is that if an invalid switch is plugged into an existing fabric it can cause significant disruptions, including modifications to zoning, and unauthorized access to fabric devices or resources.

Device-device traffic

This category includes risks coming from devices that do not talk directly to the switches, but instead use the fabric as the medium over which to talk to other connected devices, such as storage elements or hosts. The concerns here are unauthorized access and the possibility of a denial of service (DOS) attack.

The standard bodies are working on security enhancements to the Fibre Channel protocol. The vendors of Fibre Channel products will implement these enhancements using various products and features and adherence to FC security standards over time.

3.8.3 Fibre Channel security

Since April 2002, the ANSI T11 group has been working on FC-SP, a proposal for the development of a set of methods that allow security techniques to be implemented in a SAN.

Up until now, fabric access of Fibre Channel components was attended to by identification (**who are you?**). This information could be used later to decide if this device was allowed to attach to storage (by zoning), or it was just for the propagation of information (for example, attaching a switch to a switch) — but it was not a criteria to refuse an inter-switch connection.

As the fabric complexity increases, more stringent controls are required for guarding against malicious attacks and accidental configuration changes. Additionally, increasingly more in-fabric functionality is being proposed and implemented that requires a closer focus on security.

The customer demand for protecting the access to data within a fabric necessitates the standardization of interoperable security protocols. The security required within a Fibre Channel fabric to cope with attempted breaches of security can be grouped into four areas:

Authorization	I tell you what you're allowed to do!
Authentication	Tell me about yourself; I will decide if you may log in. A digital verification of who you are, it ensures that received data is from a known and trusted source.
Data confidentiality	Cryptographic protocols ensure that your data was unable to be read or otherwise utilized by any party while in transit.
Data integrity	Verification that the data you sent has not been altered or tampered with in any way.

3.8.4 Security mechanisms

In the topics that follow, we overview some of the common approaches to securing data.

Encryption

in 1976, W.Diffie and M.Hellman (their initials are found in **DH-CHAP**) introduced a new method of encryption and key management. A public-key cryptosystem is a cryptographic system that uses a pair of unique keys (a public key and a private key). Each individual is assigned a pair of these keys to encrypt and decrypt information. A message encrypted by one of these keys can only be decrypted by the other key in the pair:

- ▶ The public key is available to others for use when encrypting information that will be sent to an individual. For example, people can use a person's public key to encrypt information they want to send to that person. Similarly, people can use the user's public key to decrypt information sent by that person.
- ▶ The private key is accessible only to the individual. The individual can use the private key to decrypt any messages encrypted with the public key. Similarly, the individual can use the private key to encrypt messages, so that the messages can only be decrypted with the corresponding public key.

That means: exchanging keys is no longer a security concern. *A* has a public key and a private key. *A* can send the public key to anyone else. With that public key, *B* can encrypt data to be sent to *A*. Since the data was encrypted with *A*'s public key, *only A* can decrypt that data with his private key. If *A* wants to encrypt data to be sent to *B*, *A* needs *B*'s public key.

If *A* wants to testify that it was the person that actually sent a document, *A* will encrypt and protect the document with his private key, while others can decrypt it using *A*'s public key; they will know that in this case only *A* could have encrypted this document. Each individual involved needs their own public/private key combination.

The remaining question is: when you initially receive someone's public key for the first time, how do you know it is them? If "spoofing" someone's identity is so easy, how do you knowingly exchange public keys and how do you trust the user is who they say they are? The answer is to use a digital certificate. A digital certificate is a digital document that vouches for the identity and key ownership of an individual — it guarantees authentication and integrity.

The ability to perform switch to switch authentication in FC-SP enables a new concept in Fibre Channel: the secure fabric. Only switches that are *authorized* and properly *authenticated* are allowed to join the fabric.

Whereas, authentication in the secure fabric is twofold: the fabric wants to verify the identity of each new switch before joining the fabric, and the switch that is wanting to join the fabric wants to verify that it is connected to the right fabric. Each switch needs a list of the WWNs of the switches authorized to join the fabric, and a set of parameters that will be used to verify the identity of the other switches belonging to the fabric.

Manual configuration of such information within all the switches of the fabric is certainly possible, but not advisable in larger fabrics. And there is the need of a mechanism to manage and distribute information about authorization and authentication across the fabric.

Authorization database

The fabric authorization database is a list of the WWNs and associated information like domain-IDs of the switches that are authorized to join the fabric.

Authentication database

The fabric authentication database is a list of the set of parameters that allows the authentication of a switch within a fabric. An entry of the authentication database holds at least the switch WWN, authentication mechanism Identifier, and a list of appropriate authentication parameters.

Authentication mechanisms

In order to provide the equivalent security functions that are implemented in the LAN, the ANSI T11-group is considering a range of proposals for connection authentication and integrity which can be recognized as the FC adoption of the IP security standards. These standards propose to secure FC traffic between all FC ports and the domain controller. These are some of the methods that will be used:

- ▶ **FCPAP** refers to Secure Remote Password Protocol (SRP), RFC 2945.
- ▶ **DH-CHAP** refers to Challenge Handshake Authentication Protocol (CHAP), RFC 1994.
- ▶ **FCsec** refers to IP Security (IPsec), RFC 2406.

3.8.5 IP security

There are standards and products available originally developed for the LAN and already installed worldwide. These can easily be added into and used by SAN solutions.

Simple Network Management Protocol (SNMP) had been extended for security functions to SNMPv3. The SNMPv3 specifications were approved by the Internet Engineering Steering Group (IESG) as full Internet Standard in March 2002.

IPSec uses cryptographic techniques obtaining management data that can flow through an encrypted tunnel. Encryption makes sure that only the intended recipient can make use of it. (RFC 2401).

Other cryptographic protocols for network management are Secure Shell (SSH) and Transport Layer Security (TLS, RFC 2246). TLS was formerly known as Secure Sockets Layer (SSL). They help ensure secure remote login and other network services over insecure networks.

Remote Authentication Dial-In User Service (RADIUS) is a distributed security system developed by Lucent Technologies InterNetworking Systems. RADIUS is a common industry standard for user authentication, authorization, and accounting (RFC 2865). The RADIUS server is installed on a central computer at the customer's site. The RADIUS Network Access Server (NAS), which would be an IP-router or switch in LANs and a SAN switch in SANs, is responsible for passing user information to the RADIUS server, and then acting on the response which is returned to either permit or deny the access of a user or device.

A common method to build trusted areas in IP networks is the use of firewalls. A firewall is an agent which screens network traffic and blocks traffic it believes to be inappropriate or dangerous. You will use a firewall to filter out addresses and protocols you do not want to pass into your LAN. A firewall will protect the switches connected to the management LAN and allows only traffic from the management stations and certain protocols that you will define.



SAN disciplines

One of the key elements of a successful SAN installation is the physical location of the equipment and the disciplines introduced to manage those elements. Typically a new SAN installation is born based on an individual requirement at the time. The new SAN usually starts small and simple, but will grow very rapidly. Disciplines which were not an issue with the small SAN become major management problems as the SAN develops. It is very difficult to introduce standards to an established SAN, so careful consideration at the conceptual phase will be rewarded.

In this chapter we look at some of the SAN disciplines that should be considered prior to implementing a SAN, and look at the potential effects of not implementing these disciplines.

We will start from the floor up and make general observations and recommendations along the way to building the SAN. In the following sections we consider the pre-planning activity and compare the pros and cons of some of the options.

Remember, simply connecting the SAN components together is not a challenge, but development of a decent SAN design is one.

4.1 Floor plan

In comparison to a traditional open server environment based on SCSI technology, with Fibre Channel Protocol (FCP) we are no longer faced with short distance limitations and are able to spread our SAN over thousands of kilometers. This has its benefits for things like Disaster Recovery, but the more you distribute the SAN, the higher the cost and management overhead.

4.1.1 SAN inventory

Prior to establishing a floor plan, it is good practice to establish a high level inventory list of the SAN components that already exist, and those that will be added to the SAN. This list, which can include logical and physical components, can be used to plan the quantity and location of the SAN fabric cabinets and will feed into a more detailed list that will help design the SAN layout.

The list should include the following:

- ▶ Server type (vendor, machine type and model number)
- ▶ Switch/director type (vendor, machine type and model number)
- ▶ Storage type (vendor, machine type and model number)
- ▶ Fibre Channel protocols that devices support and cannot support
- ▶ Device (server, storage, SAN components) names and description
- ▶ Distances between devices (maximum and minimum)
- ▶ Location of admin consoles or management servers
- ▶ Storage partitioning
- ▶ Location of SCSI drives (no more than 25 m away)
- ▶ Fabric names
- ▶ Zone names
- ▶ IP addresses
- ▶ Naming conventions employed
- ▶ Passwords and userids
- ▶ Current cabinet address
- ▶ Operating systems, maintenance level and firmware levels
- ▶ Quantity and type of adapters installed
- ▶ List of WWNs and WWPNS
- ▶ If devices will have single or multiple attachments in the SAN
- ▶ Cabling cabinets
- ▶ Labels for cables
- ▶ Cable routing mapped
- ▶ Current connections
- ▶ Current configurations

4.1.2 Cable types and cable routing

There are a number of different types of cable that can be used when designing a SAN. The type of cable and route it will take all need consideration. The following section details various types of cable and issues related to the cable route.

Distance

The Fibre Channel cabling environment has many similarities to telco and open systems environments. The increase in flexibility and adaptability in the placement of the electronic network components is similar to the LAN/WAN environment, and a significant improvement over previous data center storage solutions.

Single-mode or multi-mode

Every data communications fiber belongs to one of two categories:

- ▶ Single-mode
- ▶ Multi-mode

In most cases, it is impossible to distinguish between single-mode and multi-mode fiber with the naked eye unless the manufacturer follows the color coding schemes specified by the FC-PH (see 2.7.2, “List of evolved Fibre Channel standards” on page 79) working subcommittee (typically orange for multi-mode and yellow for single-mode). There may be no difference in outward appearance, only in core size. Both fiber-optic types act as a transmission medium for light, but they have different diameters and different demands for the spectral width of the light sources:

- ▶ **Single-mode (SM):** This mode, also called mono-mode fiber, or single-mode fiber, allows for only one pathway, or propagation mode, of light to travel within the fiber. The core size is typically 8.3 - 10 μm . SM fibers are used in applications where low signal loss and high data rates are required, such as on long spans between two system or network devices, where repeater/amplifier spacing needs to be maximized. SM fiber links use longwave laser at 1270-1300 nm wavelength.
- ▶ **Multi-mode (MM):** This mode, also called multi-mode fiber, allows more than one mode of light. Common MM core sizes are 50 μm and 62.5 μm . MM fiber links can either use a shortwave (SW) laser operating at 780-860 nm, or a longwave (LW) laser at 1320-1300 nm wavelength. The low-cost shortwave laser is based on the laser diode developed for the CD players and benefits from the high volume production with that market. That makes shortwave/MM- equipment more economical. MM fiber is therefore the ideal choice for short distance applications between Fibre Channel devices.

For the supported distances of 1 Gb/s and 2 Gb/s links, refer to 2.2.1, “Small Form Factor Optical Transceivers” on page 21, and 2.2.2, “Gigabit Interface Converters” on page 23.

In Figure 4-1 we show the differences in single-mode and multi-mode fiber routes through the fiber-optic cable.

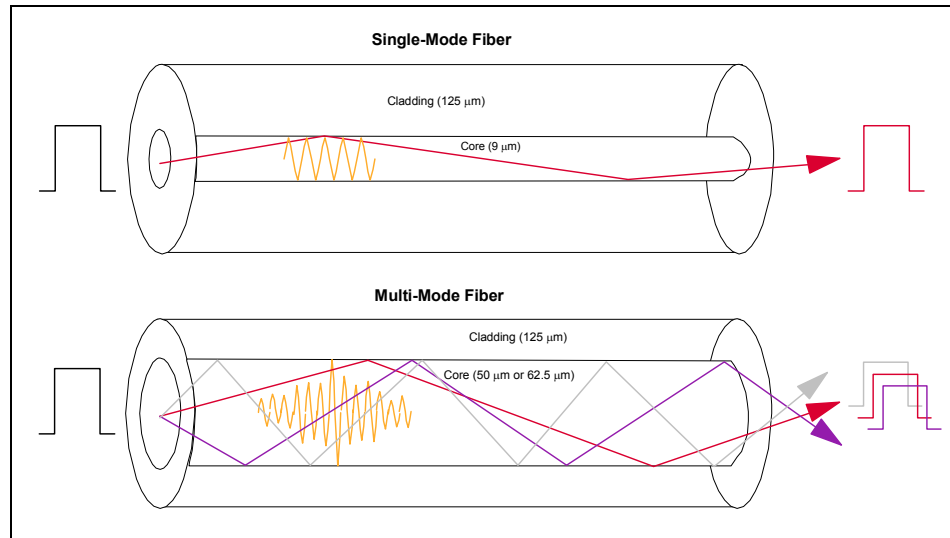


Figure 4-1 Mode differences through the fiber-optic cable

Propagation mode: The pathway of light is illustrative in defining a mode. According to electromagnetic wave theory, a mode consists of both an electric and a magnetic wave mode, which propagates through a waveguide. To transport a maximum of light, we need to have a total internal reflection on the boundary of core and cladding. With the total reflection, there comes a phase shift of the wave.

We look for modes that have the same wave amplitude and phase at each reflection to interfere constructively by wave superposition. With help of the mode equation we will find modes for a given electromagnetic wave (the light with its wavelength) propagated through a given waveguide (a fiber with its geometry). We call a waveguide *mono-mode* when only the lowest order bound mode (fundamental mode of that waveguide) can propagate.

There is exhaustive technical and scientific material about fiber and optics in “Optics2001.com”, the free Optical Community at the Web site:

<http://www.optics2001.com/Optical-directory.php>

Fiber optic cable

Fiber optic cable for telecommunications consists of three components:

- ▶ Core
- ▶ Cladding
- ▶ Coating

Core

The core is the central region of an optical fiber through which light is transmitted. In general, the telecommunications industry uses sizes from 8.3 μm to 62.5 μm . As already discussed, the standard telecommunications core sizes in use today are 8.3 (9) μm , 50 μm and 62.5 μm .

Note: Microns or micrometers (μm)? A micron is 0.0000394 (approximately 1/25,000th) of an inch, or one millionth of a meter. In industrial applications, the measurement is frequently used in precision machining. In the technology arena, however, microns are most often seen as a measurement for fiber-optic cable (which has a diameter expressed in microns), and a unit of measure in the production of microchips.

Micrometer is another name for a micron, but it is more commonly used for an instrument that measures microns in a wide variety of applications, from machine calibration to the apparent diameter of celestial objects.

Cladding

The diameter of the cladding surrounding each of these cores is 125 μm . Core sizes of 85 μm and 100 μm have been used in early applications, but are not typically used today. The core and cladding are manufactured together as a single piece of silica glass with slightly different compositions, and cannot be separated from one another.

Coating

The third section of an optical fiber is the outer protective coating. This coating is typically an ultraviolet (UV) light-cured acrylate applied during the manufacturing process to provide physical and environmental protection for the fiber. During the installation process, this coating is stripped away from the cladding to allow proper termination to an optical transmission system. The coating size can vary, but the standard sizes are 250 μm or 900 μm . The 250 μm coating takes less space in larger outdoor cables. The 900 μm coating is larger and more suitable for smaller indoor cables.

The 62.5 µm multi-mode fiber was included within the standard to accommodate older installations which had already implemented this type of fiber-optic cable. Due to the increased modal dispersion and the corresponding distance reduction of 62.5 µm multi-mode fiber, 50 µm multi-mode fiber is the preferred type for new installations. It is recommended to check with any SAN component vendor to see if 62.5 µm is supported.

For more details about fiber-optic cables, visit the “American National Standard for Telecommunications” Glossary at the Web site:

<http://www.atis.org/tg2k/tlg2k.html>

Structured and non-structured cables

In this topic we look at two types of cables: non-structured and structured.

Non-structured cables

Non-structured cables consist of a pair of optical fibers that provide two unidirectional serial bit transmission lines; they are commonly referred to as a jumper cable. Jumper cables are typically used for short links within the same room. They can be easily replaced if damaged, so they are most suited to connecting SAN components that may require regular cabling alterations.

Multi-jumper cables are available with more than one pair of fibres. They are typically used to connect more than one pair of Fibre Channel ports.

As individual cables can become easily tangled and difficult to locate, they are best avoided for longer under-floor runs.

Structured cables

Structured cables consist of multiple fiber optic cables wrapped as a single cable that have a protective member and outside jacket, and these cables are commonly referred to as trunk cables. Trunk cables are normally terminated at each end into the bottom of a patch panel. Jumper cables are then used from the top of the patch panel to the SAN fabric.

Typically trunk cables are used for longer runs between server and SAN fabric cabinets, as the trunk cable terminates at a patch panel, normally there is no requirement to make future cable alterations.

In Table 4-1 we compare the advantages and disadvantages of using non-structured and structured cabling practices for server cabinet to SAN fabric cabinet connections.

Table 4-1 Comparison between structured and non-structured cables

Non-structured cables	Structured cables
Unknown cable routing	Known cable route
No cable documentation system	Defined cable documentation
Unpredictable impact of moves, adds and changes	Reliable outcome of moves, adds and changes
Every under floor activity is a risk	Under floor activity can be planned to minimize risk
No waste, only the cables required are run	Initially not all fibers will be used; spare cables will be run for growth

There are a number of companies that provide fiber cabling service options. IBM provide this service offering with Fibre Transport Service Cabling System (FTS).

When planning the use of trunk cable over a longer distance it is important to consider the potential light loss. Every time a joint is made in a fiber cable there will be a slight light loss, with the termination at the patch panel there will be considerably more light loss.

The fibre installation provider should be able to calculate potential light loss to ensure the trunk cable run is within acceptable light loss limits.

Patch panels

Patch panels are commonly used to connect trunk cables, particularly between floors and buildings. They provide the flexibility to enable repatching of fibres but generally the panel configuration will remain the same after installation. Pay attention not to mix fibre cables with different diameters when crossing patch panels.

4.1.3 Planning considerations and recommendations

Many miscellaneous considerations are needed to successfully install fiber-optic links for any protocol. However, the higher data rate and lower optical link budgets of Fibre Channel lends itself to more conservative approaches to link design. Some of the key elements to consider are:

- ▶ All links must use the currently predominant “physical contact” connectors for smaller losses, better back reflectance, and more repeatable performance.
- ▶ The use of either fusion or mechanical splices is left to the designer to determine the desired losses weighed against the cost of installation.

- ▶ Multi-mode links cannot contain mixed fiber diameters (62.5 and 50 micron) in the same link. The losses due to the mismatch may be as much as 4.8 dB with a variance of 0.12 dB. This would more than exceed the small power budgets available by this standard.
- ▶ The use of high quality factory terminated jumper cables is also recommended to ensure consistent performance and loss characteristics throughout the installation.
- ▶ The use of a structured cabling system is strongly recommended even for small installations.
- ▶ A structured cabling system provides a protected solution that serves current requirements as well as allows for easy expansion.
- ▶ The designer of a structured system should consider component variance affects on the link if applicable.

Much of the discussion so far has been centered around single floor or single room installation. Unlike earlier FDDI or ESCON installations that had sufficient multi-mode link budgets to span significant distances, Fibre Channel multi-mode solutions for the most part do not. Though the Fibre Channel standard allows for extend distance links and handles distance timing issues in the protocol the link budgets are the limiting factor.

Therefore, installations that need to span between floors or buildings will need any proposed link to be evaluated for its link budget closely. Degradation over time, environmental effects on cables run in unconditioned spaces, as well as variations introduced by multiple installers need to be closely scrutinized. The choice between single-mode and multi-mode devices may need to be made for many more links. Repeating the signal may also provide a cost effective solution if intermediary conditioned space can be found.

Since Fibre Channel provides a built in mirroring capability to SAN, in addition to its 10 km link distances using single-mode fiber, there will be more consideration for off-campus or across city links. In these cases, right-of-way issues, leasing of “dark” fiber (no powered devices provided by the lessors) issues, service level agreements, and other factors associated with leaving the client owned premises needs to be planned for and negotiated with local providers. The industry has also announced interest in providing wide area network (WAN) interfaces similar to those employed in the networking world of today. When these devices are made available, then connections to these devices will need to be included in the designs as well.

4.1.4 Structured cabling

Because of access to the Internet, the data centers of today are changing rapidly. Both e-business and e-commerce are placing increasing demands on access to and reliance on the data center. No longer is the data center insulated from the rest of the company and just used to perform batch processing.

Now, access and processing is a 24x7 necessity for both the company and its customers. The cabling that connects servers to the data storage devices has become a vital part of corporate success. Few companies can function without a computer installation supported by an efficiently structured and managed cabling system.

There are many important factors to consider when planning and implementing a computer data center. Often, the actual physical cabling is not given enough planning and is considered only when the equipment arrives. The result of this poor planning is cabling that is hard to manage when it comes to future moves, adds, and changes due to equipment growth and changes.

Planning a manageable cabling system requires knowledge about the equipment being connected, the floor layout of the data center(s), and, most importantly, how the system requirements will change. Questions that should be considered include:

- ▶ Will the data center grow every year?
- ▶ Will you need to move the equipment around the floor(s)?
- ▶ Will you upgrade the equipment?
- ▶ Will you add new equipment?
- ▶ What type of cabling do you require?
- ▶ How will you run the cables?
- ▶ How will you label the cables?
- ▶ Can you easily trace the cables if there is a problem?

Answers to these important questions should be obtained as part of the early planning for the cabling installation.

4.1.5 Data center fiber cabling options

The most prevalent data center connectivity environment that uses fiber cabling is IBM's ESCON architecture. However, the same structured fiber cabling principles can be applied in the SAN environment, and to other fiber connectivity environments such as IBM's Fiber Connection (FICON), Parallel Sysplex, and Open Systems Adapters (OSA). The examples throughout this chapter apply to structured fiber optic cabling systems designed to support multiple fiber-optic connectivity environments.

The need for data center fiber cabling implementation arises from the following three scenarios:

- ▶ Establishing a new data center
- ▶ Upgrading an existing data center by replacing the cabling
- ▶ Adding new equipment to an existing data center

IBM can help you design and implement a network that leverages existing investments, avoids costly downtime, and saves time and money when moving to performance enhancing technologies.

IBM Network Integration and Deployment Services

IBM Network Integration and Deployment Services helps businesses integrate and deploy a complex network infrastructure that leverages multivendor technologies. IBM will analyze existing networks, protocols, wired and wireless configurations to identify performance, interoperability, and connectivity requirements. Implementation planning, detailed logical and physical network design, rapid deployment and network rollouts, product installation and customization and operational services for network and cabling infrastructures is provided.

This enables business to securely converge data, voice, and video networks, enables intelligent network infrastructures, and deploy mobility solutions by exploiting technologies such as virtual private networking (VPN), video and voice over IP (VoIP), fiber optic networking, content delivery networks, storage networking and wireless.

More information is available at Web site:

<http://www-1.ibm.com/services/networking/integration/index.html>

IBM Cabling Services and the Advanced Connectivity System

Today's telecommunications infrastructure, both copper and optical fiber, supports data rates that were undreamed of even a decade ago. The cabling infrastructure is at the core of every voice, data and multimedia network.

Integrating multivendor equipment has become challenging, time-consuming and increasingly dependent on how IT systems are physically connected. Proper planning, configuration and installation for connectivity is critical. Our professionals can analyze your existing network, protocols, wiring configurations and cabling infrastructure, identify system interoperability and connectivity requirements, and help you connect and integrate your cabling systems.

IBM has a wide array of premises, data center, server and storage networking solutions to help select and install the right cabling solution for the e-business infrastructure.

Cabling solutions using the IBM Advanced Connectivity System include copper and fiber solutions for any building or premises, and each is designed to be intermixed and adapted to different topologies. There is a choice of several grades of cabling infrastructures that meet or significantly exceed today's technical standards and performance demands.

<http://www-1.ibm.com/services/networking/integration/acs.html>

Metropolitan Area Network cables

Metropolitan Area Network (MAN) cables are typically used for business continuance between two sites. MANs used for business continuance normally consist of a diverse route of a primary and alternate cable. The alternate route is normally only used when the primary route is not available. To ensure we introduce no single points of failure it is critical these cables enter and leave the building at separate locations and at no point share the same cable run or equipment.

It is important to have a detailed intersite cable route plan to highlight any single points of failure and to determine the exact distance of both routes.

If the primary site is several kilometers shorter than the secondary route, there may be latency issues to consider when using the secondary route. It will only introduce problems, when there is parallel MAN links used in a shared manner, and the skew which comes with the different latencies cannot be handled by the protocol.

As the MAN cables enter the buildings, the routes of the primary and alternate cables should be clearly marked on the floor plan.

Ethernet cables

The majority of SAN products require IP addresses to enable remote software management. To ease the administration of SAN management it is common to place the Ethernet ports within the same LAN or VLAN and choose IP addresses from the same IP subnet.

Ethernet cables will need to be laid from the site Ethernet switch to the SAN fabric cabinet, and these cable routes should be detailed on the plan.

Future growth

Future technologies and design issues are mostly affected by length and attenuation due to increased speeds. Since future technologies are unknown, most organizations are pulling single-mode fiber along with the new multi-mode fiber while keeping the proposed distance limitations in mind when designing the cable plant.

One option is to leave the single-mode fiber un-terminated and dark for future technologies. Connectors, panels, and the labor to terminate, test, and install these items could be a significant cost, so leaving these cables un-terminated and dark can save money in the short term.

Another option is to proceed with the termination in anticipation of rapid technological developments. Indications point to a reduction in the cost of LW lasers. This would drive down the price of LW technology and LW equipment applications, influencing the adoption of single-mode usage as well.

4.1.6 Cabinets

For IBM SAN fabric components that do not come with an associated cabinet, you will have a choice of rack mount or non-rack mount feature codes. In most cases it is advisable to select the rack mount option for the following reasons.

- ▶ **Security:** To prevent unauthorized actions on the SAN fabric components, the cabinet can be locked and key access restricted to selected personnel.
- ▶ **Audit trail:** Hardware changes can be tracked by recording who has requested the cabinet key, and at what time.
- ▶ **Cable Management:** When there are large numbers of fibre cables hanging from the SAN fabric, it can be very difficult to locate and alter cables. You also have an associated risk that when you are making SAN cable alterations, you may damage other cables in the SAN fabric. The use of cable supports, cable ties or velcro strips will enable cables to be tied back along the cabinet edges, reducing the risk of accidentally damage and enabling cables to be easily identified.
- ▶ **Hardware replacement:** When customer engineers need to repair or replace a part of the SAN fabric, it is important they have easy access to the component. The use of racks will guarantee they are able to access the device without disturbing any other SAN components.
- ▶ **Power outlets:** Most cabinets have a default number of power outlets, this number can be used to plan current and future SAN fabric power requirements.
- ▶ **Component location:** Each cabinet should be clearly labelled, these labels can be incorporated within the SAN fabric components naming standards. In the event of a problem or change the correct component can be easily identified.

In Figure 4-2 we show an example of a SAN fabric that has not been racked, and where only one cable has been labelled. It is easy to see how cables could become damaged and mistakes could occur.

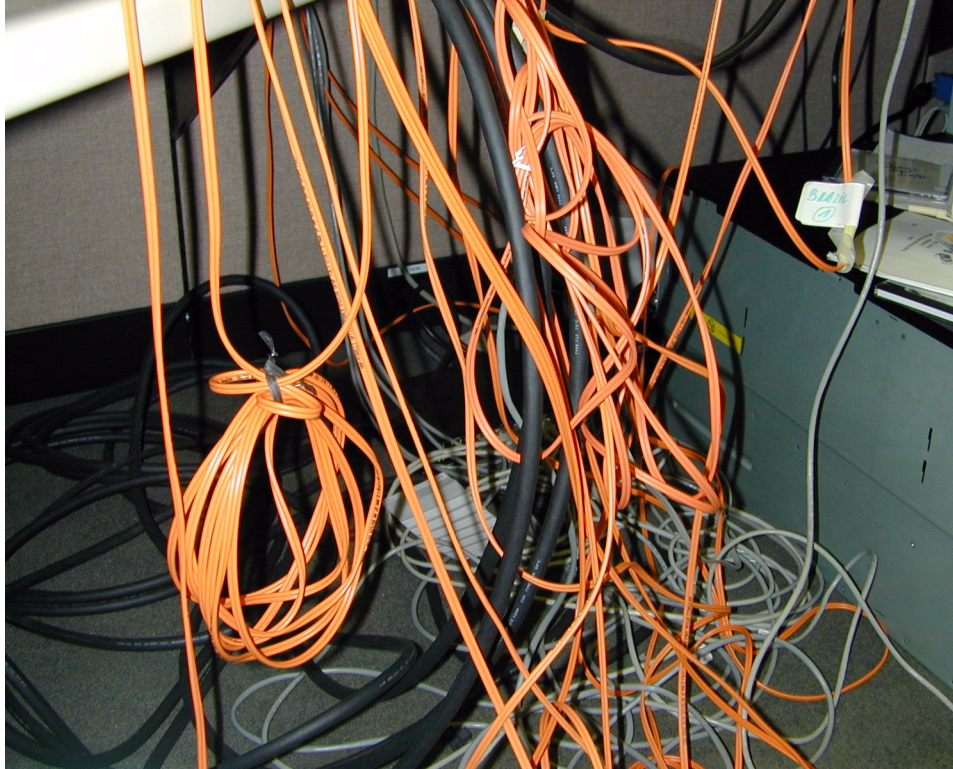


Figure 4-2 Messy cabling, no cabinet, and no cable labels

4.1.7 Phone sockets

Most of the larger SAN devices will have dial home facilities which will require a phone line. Although phone lines can be shared between devices, sufficient phone sockets need to be provided to prevent phone line bottlenecks. It is also wise to think about ensuring that spare phone lines are available should one fail, or be in use for a long period of time for any reason.

As an example, if a phone line was shared between an IBM ESS and a fabric component, and if log information had to be extracted from the ESS, the phone line could be busy for over one hour — and any potential error on the fabric component may go unnoticed.

4.1.8 Environmental considerations

In this section we consider some basic requirements for power sources and heat dissipation.

Power

The majority of SAN components have the option of single or dual power supplies. To realize the benefits of two power supplies, it is essential the power source supplying the devices is from two independent supplies.

In a cabinet full of smaller SAN fabric devices with dual power supplies, it is very easy to exceed the available number of power sockets.

You need to ensure the SAN fabric cabinet has sufficient power sockets to satisfy the SAN fabrics power requirement from both day one and a potentially full cabinet.

Heat

Several small switches in a SAN cabinet will generate quite a lot of heat. To avoid heat damage it is important the cabinet is located in a room that has temperature control facilities.

4.1.9 Location

Typically the location of the SAN fabric cabinets, servers and storage (disk and tape) will be dictated by available space and power supply.

The typical placement for components is in clusters around the periphery of the work area. This minimizes the length of cables and their exposure to points of failure. An alternative is to locate components in a central cluster. However, the smaller the area that the components are gathered in, the more potential exists for a burst pipe, for example, taking out the whole fabric. It may be wise to distance some components apart from each other.

The further that the two components are apart, the less likely it is that a single disaster will render both of them unusable.

4.1.10 Sequence for design

Assuming you are cabling a facility with existing components, the usual sequence is to do the following:

- ▶ Base it on the server inventory and detail the current components accurately and completely.
- ▶ Determine what new components will be added and their location.
- ▶ Verify that the type of cable is appropriate for each connection.
- ▶ Calculate loss and attenuation for each connection and for the total system.
- ▶ Modify the design as needed.

A detailed floorplan should be drawn with cabinet and slot locations of all components of the SAN. The floorplan should include:

- ▶ Servers
- ▶ SAN fabric
- ▶ Storage devices
- ▶ Cable routes
- ▶ Cable type
- ▶ Cable entry and exit points
- ▶ Power points
- ▶ Power source
- ▶ Phone lines
- ▶ Ethernet cable routes
- ▶ Location of SAN ethernet hub and any required ethernet switches

If two buildings are connected using a MAN or similar, the cable routes, the total distance of both the primary and secondary routes, and the entry and exit points into the building need to be detailed.

On completion of the floor plan, the checklist displayed in Table 4-2 in should be performed to validate the proposed layout.

Table 4-2 Checklist for proposed layout

Check	Validate	Successful
Location of SCSI devices	Within 25 m of SAN bridge device	
Multi-mode, shortwave devices	Within 550 m (1 Gb/s) and 300 m (2 Gb/s)	
	Cable route uses 50/125 multi-mode fiber	
Single-mode, shortwave devices	Within 275 m (1 Gb/s) and 150 m (2 Gb/s)	
	Cable route uses 62.5/125 multi-mode fiber	
Long-wave devices no extender	Within 10 km	
Long-wave devices with extenders	Within 100 km	
Power source	Independent supply	

Check	Validate	Successful
Number of power sockets	Sufficient number, including units with dual power supply	
Location of SAN devices	In a lockable cabinet	
	Sufficient space in SAN cabinets	
LAN	Sufficient free IP addresses and Ethernet ports for SAN devices	
Capacity	No physical constraints for growth - cables, full cabinets, and so on	
Phone sockets	Location and quantity of phone sockets that require dial home	

4.2 Naming conventions

Use of descriptive naming conventions is one of the most important factors in a successful SAN. Good naming standards will improve problem diagnostics, reduce human error, allow for the creation of detailed documentation and reduce the dependency on individuals.

4.2.1 Servers

Typically, servers will already have some form of naming standard in place. If a server name does exist, it should have been captured during inventory, as described in 4.1.1, “SAN inventory” on page 124.

The local server name is typically used as the host name defined to the disk system. For the ESS you would normally use the server name in the server description field. The same local server name can be used within the switched fabric as an alias for zone settings, and whenever possible the use of the server name should be consistent throughout the SAN.

4.2.2 Cabinets

SAN fabric cabinets should be labelled to adhere with local site standards.

4.2.3 SAN fabric components

A good naming convention for the SAN fabric component should be able to tell you the physical location, component type, have a unique identifier and give a description of what it connects to. The following are some descriptor fields that may be considered when designing a fabric naming convention. If your SAN only has one vendor type or only one cabinet the name could be a lot simpler.

Component description

This should describe the fabric component and the product vendor (for mixed vendor environments) which will help you locate the management interface and the component number within the SAN. For example, to give it a unique identifier you may want to use something similar to the following:

- ▶ Type — Switch (S) Director (D) Gateway (G) Hub (H) Router (R)
- ▶ Vendor — Brocade (B) Cisco (C) McDATA (M) Vicom (V)
- ▶ Number — 1 - 99

For example, the third Brocade Switch in cabinet one would be:

- ▶ S3 B

Connection description

This should detail what the component is connecting to. For highly available devices such as the ESS, it is important to understand which cluster side of the device the component is connected to. This will help prevent potential mistakes in the SAN design. For devices used to expand the SAN that do not connect to disk or tape, we will simply identify them as cascade.

- ▶ Connection — Disk (D (for ESS either cluster A or B)), Tape (T), Cascade (C)
- ▶ Number — 1 - 99

To continue our example, the third Brocade Switch in cabinet one connecting to ESS3 Cluster A would be

- ▶ S3 B D3A

Physical location

This may be the cabinet descriptor field and, for example, SAN cabinet one could be C1. For our example this would give us:

- ▶ S3 B D3A C1

We show how our name is developed in Figure 4-3.

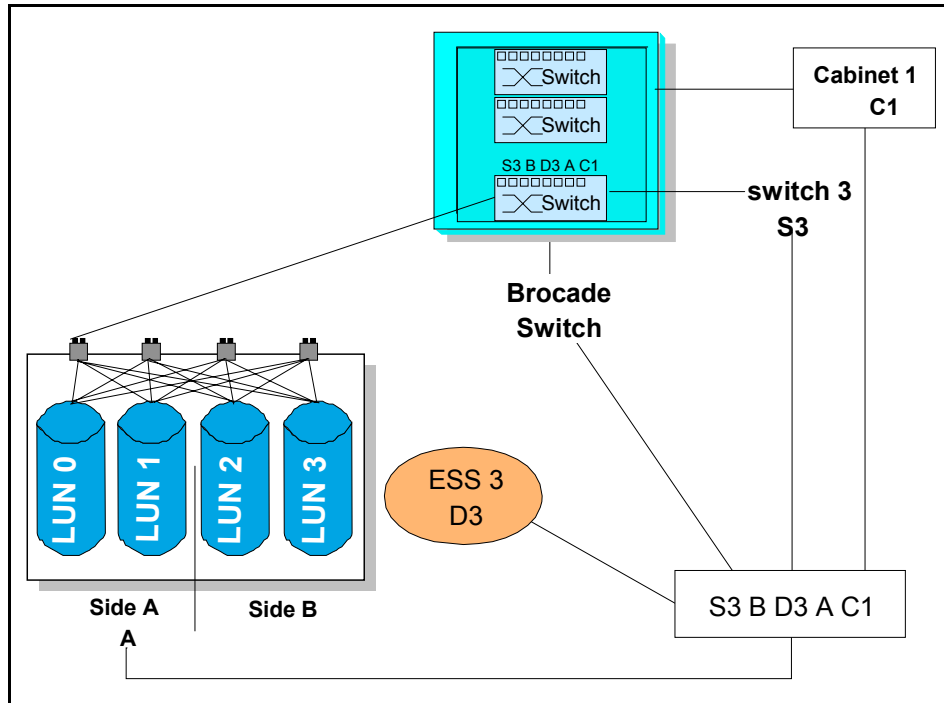


Figure 4-3 Naming convention development

4.2.4 Cables

Modifications to a SAN that does not have sufficient labelling in place could lead to the incorrect selection and reconfiguration of the SAN with potentially disastrous effects.

When determining the exact cable identification tag, try to avoid using device specific names. Such names do not take into account adding or subtracting devices, or devices being renamed, which are fundamental parts of a SAN.

The chosen cable tag naming standard should be incorporated in a detailed SAN fabric port layout plan. The port layout plan will enable you to identify the exact devices to which the cable is connected.

Any SAN cable reconfigurations should have an associated change record. Part of that change process should include a pointer to update the cable tag and port layout plan. Adhering to this plan will ensure that the document is always kept up-to-date.

4.2.5 Zones

Understanding software zones defined by another SAN administrator, when no naming standards have been defined, can be very difficult. Researching the zone setting can be time consuming and can cause problems if activated incorrectly. The introduction of site standard software zone naming standards will minimize this risk.

A good zone naming standard should consist of a meaningful description of the servers effected by the zone and what you are hoping to achieve.

For example, if we assume that the WWNs of servers ABC_FIN1 and ABC_FIN2 (the company production finance server) have been zoned to only see the WWN of disk ESS3.

If we called this zone1, the creator would probably understand what the zone was doing. If, at a later stage, the finance department requested the ability to access archive files that resided on ESS1, a modification to zone1 would be required. However, as zone1 gives us no descriptive information as to the contents of the zone itself, for anyone other than its creator, this change would be very difficult and may involve investigating every zone to see its contents.

If, however, we had called the zone PRODFINESS3, locating and modifying the zone would be much simpler.

4.3 Documentation

There are a number of software tools (such as Tivoli's TSNM) that are able to provide detailed information and documentation about the SAN. This includes connection diagrams, server utilization reports and status monitors and more besides.

These products, although very good at giving you an overall picture of the SAN, do not have sufficient detail to be the only source of information in order to manage the SAN.

Data that needs to be collected and recorded in SAN documents include the following:

- ▶ **Floorplan:** The floor plans of all SAN machine rooms
- ▶ **Server Inventory:** A list of servers connected to the SAN, type of Host Bus Adapters (HBAs), World Wide Name of HBAs
- ▶ **List of fabric components:** The naming convention and list of all fabric components

- ▶ **Space allocated:** A list of LUNs allocated to servers
- ▶ **Space available:** A list of free space in the disk device
- ▶ **Fabric Connection:** A detailed wiring diagram of the SAN fabric
- ▶ **Fabric Port layout:** A port usage plan detailing what ports are currently used, which ports are spare
- ▶ **Zone Information:** Both hard and soft zoning in place
- ▶ **IP addresses:** A list of IP addresses for all fabric components, as well as a list of spare ones
- ▶ **Fabric model Serial numbers:** The IBM product serial numbers (used when raising a call with the IBM call center)
- ▶ **Micro code versions:** The level of micro code installed on the disk devices (used when raising a call with IBM call center)
- ▶ **Firmware version:** The level of firmware running on the SAN fabric (used when raising a call with IBM call center)
- ▶ **Procedures:** A step-by-step how to perform a SAN function guide

In addition to the information documented for the primary site, there will also be a requirement for a similar level of documentation for the disaster site.

4.4 Power-on sequence

After a site power-down, it is important to stagger the power-up of the servers connected to the SAN fabric. The reason for this is that during boot-up some operating systems will scan all the switch ports and will look up other HBA ports.

With some combinations of HBA cards this can have an adverse effect on other servers in the SAN. Symptoms can be unpredictable, ranging from clusters being brought down, to NT losing SAN access and requiring a reboot.

The use of soft zoning to separate vendor cards would prevent the risk of this occurring.

4.5 Security

Consolidating storage onto central devices has many benefits, but can also increase the risks to your business. With large amounts of critical data in one location, it is important to ensure that you are providing the maximum protection of your data. That topic was discussed from a general point of view in 3.8, “Security” on page 115.

4.5.1 General

All SAN software management tools come with a default userid and password which typically has the highest level of authority. Obtaining unauthorized access to these IDs would enable a user to alter zone information and give servers access to data that would otherwise be protected.

Generally, SAN software products do not police their userids and passwords and will not request them to be changed. It is common to find default IDs remaining on the system months after the SAN has been installed. The userids and passwords need to be changed as part of the installation, and passwords should be altered at regular intervals from then on.

Operational controls

In 3.8.1, “Control types” on page 116, operational controls were described as part of a whole security concept. Numerous attempts are required to structure and define single steps and tasks in order to archive the highest possible security level in the IT environment. Such tasks as backup and recovery, physical security, and so on, are defined in policies and grouped in operational controls.

The Acceptable Use Policy defines acceptable use of equipment and computing services, and the appropriate employee security measures to protect the organization’s corporate resources and proprietary information.

A security policy can start simply as an Acceptable Use Policy for network resources, and may grow to large documents as a complete set of laws, rules, and practices that regulate how an organization manages, protects, and distributes sensitive information. In such a policy, you can state which administration group will have access to which components: Does your SAN administration manage your servers too? (see 7.1, “Overview” on page 256 about that topic). By building up your security policy, you will define and publish your security rules. RFC 2196 suitably defines a security policy in a 73-page *Site Security Handbook*.

RFC 2196: “A security policy is a formal statement of the rules by which people who are given access to an organization's technology and information assets must abide.”

The full text of RFC 2196 is available under this URL:

<http://www.ietf.org/rfc/rfc2196.txt>

4.5.2 Physical access

Physical security is an absolutely essential component of any comprehensive security plan. Even with excellent software controls in place, physical access to enterprise elements opens the door to a whole range of security issues. To ensure physical security, fabric devices should reside in environments where physical access controls provide adequate protection.

Secure machine room

With the flexibility of a SAN there is the temptation to distribute the SAN fabric in the location of the servers. This should be avoided if the locations cannot be adequately protected.

Cabinet protection

As detailed in 4.1.6, “Cabinets” on page 134, fabric cabinets should be lockable with restricted access to the key.

Switch protection

SAN switches usually provide RS-232 and Ethernet connections. Access to either of these interfaces must only be given to trusted persons, as all of the vital data of switches and fabric can be monitored and changed from here.

Cable protection

Damage to a fiber optic cable can result in performance degradation or a complete loss of access to the data. Fibers should be laid in cable trays or trunks with rodent control measures in place.

4.5.3 Remote access

There are a variety of ways to obtain information from fabric switches. Common management access methods involve the use of telnet for command line functionality, HTTP for Web-based access, in-band Fiber Channel for management server access, and console access for direct switch connectivity. Common to all of these applications is that they need IP connectivity — and the IT community has been alarmed for years about how many ways there are to break into IP hosts. Each of the possible access methods has its associated security issues.

Telnet: The essential problem with telnet access is that it transmits unencrypted the username, password, and all data going between the management system and the switch. Any user with a promiscuous network interface card and data-sniffing programs can capture the whole data transfer back and forth, including account and password.

HTTP: Similar to the telnet issues mentioned above, when a system uses a Web-based application like Web Tools to logon (authenticate) to the switch in order to run privileged commands, it passes the login information not encrypted.

Management Server: This remote management method uses an in-band Fibre Channel connection to administer or obtain information from the fabric switches. By default, it grants access to any device. However, it is possible to create an access control list to limit the WWNs of devices that can connect to the switch using this method.

Console Access: Although not usually thought of for remote access, it is possible to adapt console connections to remote use through the use of terminal server devices. Thus, an organization can use telnet, secure shell (SSH), or some similar application to connect to the terminal server, which then in turn connects to the selected device through the console interface. This solution has the potential to provide additional security through the use of third-party products.

You'll find customers who lack a firm security policy and so leave userids and passwords as default, and apply public network addresses to their switches — and then there is the other extreme: customers who disconnect the switches from the IP network. To satisfy the needs of security *and* manageability, a SAN needs the IP-connectivity *and* the strongest possible security features available.

4.6 Education

It is important that the educational requirements of all those involved in implementing and maintaining the SAN are considered in order to gain the maximum benefit from the SAN — and minimize the room of remaining human error. Therefore, the right skills have to be defined and a way to validate these skills.

4.6.1 SAN administrators

The SAN administrator is commonly responsible for effective utilization of the SAN resource, resource protection, balancing traffic, performance monitoring, utilization trending, and error diagnostics, in addition to many maintenance functions. The SAN administrator must be identified as the focal point for any additions, deletions or modifications of the SAN environment.

Management of the SAN is usually performed using the software interface that comes with each of the SAN fabric components. There are a number of software products that enable all components of the SAN to be monitored and managed from a central point. Most SAN software management tools have facilities to create different levels of access and these range from view through to full administration.

4.6.2 Skills

As already stated in 3.8, “Security” on page 115, the Fibre Channel fabric and its components are considered shared resources. This accessibility, plus the increase in both quantity of data and users of that data, has heightened the awareness of the IT community to security exposures. As such, Fibre Channel fabrics and their components are prone to the same types of security breaches once associated only with IP networks. These concerns of the SAN community reflect the same concerns of the IT community.

But not only is security a major link between SANs and LANs — the nature of connectivity is the same and it is likely to become even closer as technology matures.

Good networking skills are needed to implement and operate SANs — the kind of skills which have typically been developed in LAN environments. Good operating system and diverse platform skills are also required with this increase in connectivity — so both mainframe and open systems skills may now be needed in the same person.

The SCSI heritage in the SAN is enormous. Good storage skills must not be overlooked. The combination for SAN success in implementation and services will likely be a merged skill of networking, storage, system, and security skills.

4.6.3 Certification

Although not a part of education, certification is a good indicator as to the core competency and ability of an individual. There are a number of programs available.

IBM Professional Certification Program

The IBM Professional Certification Program is designed to validate technical skills to network administrators and integrators, systems integrators, solution architects and developers, resellers, technical coordinators, sales representative, or educational trainers.

The Program has developed certification role names to guide the participants in their professional development. The certification role names include IBM Certified Specialist, IBM Certified Solutions/Systems Expert, and IBM Certified Advanced Technical Expert for technical professionals who sell, service, and support IBM solutions. For technical professionals in application development, the certification roles include IBM Certified Developer Associate and IBM Certified Developer. IBM Certified Instructor certifies the professional instructor.

The Professional Certification Program from IBM provides with a structured program leading to an internationally recognized qualification.

Among other programs such as AIX, Linux, DB2®, WebSphere®, Lotus®, and Tivoli, IBM Professional Certification Program provides SAN certifications by offering Enterprise Tape and Disk Solutions, Open Systems Storage, and as a pure SAN certification: TotalStorage Networking Solutions.

IBM TotalStorage Networking Solutions

The IBM Certified Specialist designs IBM TotalStorage end-to-end storage networking solutions to meet customer needs. This individual provides comprehensive storage networking solutions that include servers, storage networking, storage devices, management software, and services. This specialist has detailed knowledge of SAN, NAS, and iSCSI technologies and the corresponding management software. He or she has broad knowledge of IBM storage products and their features and functions, and can describe in detail the storage networking strategy and solutions, industry, competition, and business trends.

To learn more about the IBM Professional Certification Program, visit the Web site:

<http://www-1.ibm.com/certify/index.shtml>

SNIA Storage Networking Certification Program

The Storage Networking Industry Association (SNIA) is introducing the industry's first vendor-independent certification program for storage networking called SNCP. The program was developed in response to demand from enterprise customers worldwide in order to provide standards for measuring the storage networking expertise of IT professionals.

The SNIA is identifying the technologies that are integral for IT professionals to understand and deploy storage networks. The first modules of the SNIA SNCP, developed for the SNIA by the industry-leading training company Infinity I/O, include certification exams testing candidates' knowledge of Fibre Channel SANs. Future modules of the SNIA SNCP are expected to include storage networking topics such as NAS and IP Storage, as well as applications such as backup and restore and capacity planning.

SNCP currently offers three levels of certification:

- ▶ Level 1 - Fibre Channel Storage Networking Professional
- ▶ Level 2 - Fibre Channel Storage Networking Practitioner
- ▶ Level 3 - Fibre Channel Storage Networking Specialist

To learn more about SNCP and the depth of the different certification levels, visit the Web site:

<http://www.snia.org/education/certification/>



Host Bus Adapters

The IBM supported SAN environments contain a growing selection of server Fibre Channel Host Bus Adapters (HBAs), each with their own functions and features. For the majority of open systems platforms, this presents us with the opportunity to select the most suitable card to meet the requirements of the SAN design.

In this chapter we provide an overview of the IBM supported HBAs and highlight any unique functions the particular card may have.

Note: For some open systems platforms, the supported HBA is actually provided by the vendor. In most cases the HBAs used by the vendor are manufactured by one of the main HBA providers detailed in this section. For example, the HBA FC 6227 supported for pSeries servers is supplied by Emulex.

Nevertheless, this chapter can offer some value to readers of these platforms, as we provide an overview of each HBA and give detailed error diagnostic tips that would still apply to these platforms.

5.1 Selection criterion

In this section we look at a number of points that should be considered when selecting the right HBA to meet your requirements.

5.1.1 IBM supported HBAs

The first and most important factor to consider, when selecting a Fibre Channel HBA, is whether it is supported by IBM for the server make and model, and also the manner in which you intend to implement the server. For example, an HBA may be supported for the required server, but if you require dual pathing or the server to be clustered, the same HBA may no longer be supported.

To ensure that the HBA is supported by IBM in the configuration you require, refer to:

<http://www.storage.ibm.com/disk/ess/supserver.htm>

For IBMers only, for an HBA that is not detailed as supported for a specific platform, support can be requested using the Request Product Quotation (RPQ) process.

5.1.2 IBM SSG HBA and SAN interoperability matrix

For a list of the currently supported IBM SAN and storage components, refer to the Web site:

http://ssddom02.storage.ibm.com/hba/hba_support.pdf

5.1.3 ESS host systems attachment

For detailed instructions of how to connect HBAs to the ESS, refer to the manual:

- ▶ *IBM Enterprise Storage Server Host Systems Attachment Guide, 2105 Models E10, E20, F10 and F20, SC26-7296*

5.1.4 Special features

Any special functions you require from your SAN need to be considered, as not all HBAs may support the function. These functions could include:

- ▶ Dual connection
- ▶ Performing an external server boot
- ▶ Connection to mixed storage vendors
- ▶ Fault diagnostics

5.1.5 Quantity of servers

Another factor to consider is the number of servers in your environment that will require Fibre Channel HBAs. Having a common set of HBAs throughout your SAN environment has a number of advantages:

- ▶ It is easier to maintain the same level of firmware for all HBAs.
- ▶ The process for downloading and updating firmware will be consistent.
- ▶ Firmware and device driver can be a site standard.
- ▶ Any special BIOS settings can be a site standard.
- ▶ Fault diagnostics will be consistent.
- ▶ Error support will be from a single vendor.

5.1.6 Product specifics

In the topics that follow, we look at three vendors that are associated with the IBM portfolio of HBAs.

5.2 Emulex

At the time of writing, IBM currently supports the Emulex LP8000, LP8000S, LP9002DC, LP9002L, and LP9002S Fibre Channel adapters. Refer to the HBA interoperability matrix to find the version that is supported for your operating system at the Web site:

<http://www.emulex.com/ts/docoem/framibm.htm>

5.2.1 LP7000E

The Light Pulse LP7000E, a second generation Fibre Channel PCI host bus adapter, uses the Emulex Superfly chipset, a 266 MIPS onboard processor and high speed buffer memory. The LP7000E features a 32-bit PCI interface.

The 32-Bit memory acts as a frame buffer and enables the LP7000 to achieve its high performance throughput.

The 1 Gb/s LP7000E provides features, including switched fabric support using F_Port and FL_Port connections, full-duplex data transfers, high data integrity features, support for all Fibre Channel topologies, and support for service classes 2 and 3.

5.2.2 LP8000

In Figure 5-1 we show the third generation Fibre Channel PCI host bus adapter, the LP8000 which uses the Dragonfly ASIC with a 266 MIPs onboard processor and offers 128 KB buffer RAM. It supports simultaneous full duplex 1 Gb/s, which delivers up to 200 MB/s.

Similar to the 64 buffer credit associated with a longwave port in a switch or director, the 64 Bit interface enables the LP8000 to sustain high performance over a distance of up to 10 km. This buffer capability improves the performance of the card.

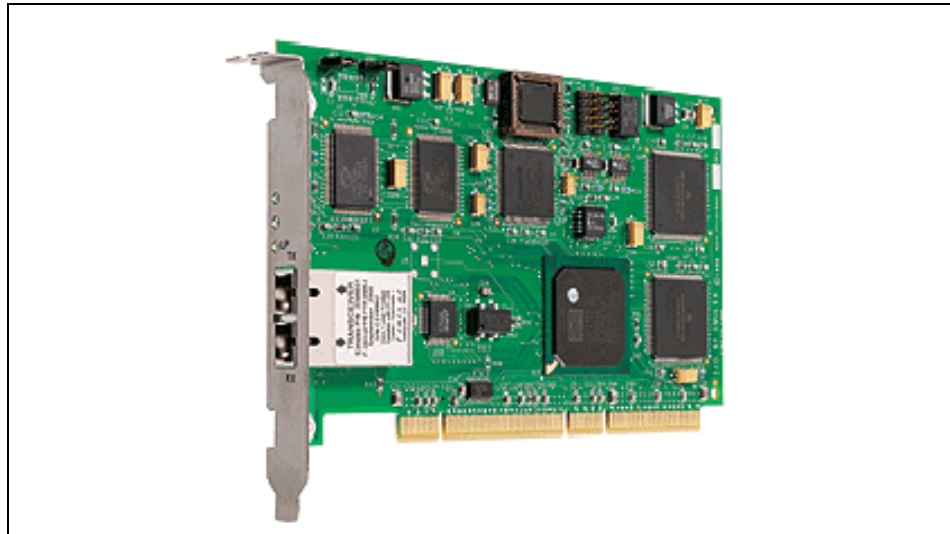


Figure 5-1 Emulex LP8000 HBA

The LP8000 also features sophisticated hardware that provides superior performance in storage area networks, delivering low latency and high throughput in switched, arbitrated loop, and clustered environments. Support for both copper and fiber optic cabling is provided through standard GBIC or embedded interfaces.

5.2.3 LP8000S

The LP8000S is a 64-bit SBus Fibre Channel HBA which has been optimized for Sun Microsystems SBus servers. It features the same 266 MIP Dragonfly ASIC as the LP8000 and offers 128 KB buffer RAM. It also supports simultaneous full duplex, which delivers up to 200 MB/s. The LP8000S features common hardware/firmware as the LP8000. Supported distances are up to 500 m, depending on the fiber optic cabling being used.

5.2.4 LP9002DC

The LP9002DC Dual Channel PCI host bus adapter offers two independent 2 Gb/s Fibre Channel HBA interfaces in a single PCI slot. It features two Emulex Centaur ASICs, two 266MIPS onboard processors, 128 KB RAM, and a high performance 64-bit 66 MHz PCI bridge.

The LP9002DC features an automatic speed negotiation capability that allows complete compatibility with existing 1 Gb/s Fibre Channel SANs, while allowing seamless upgrades to higher speed 2 Gb/s SANs. The LP9002DC architecture is based on two of Emulex's high performance 2 Gb/s HBAs integrated into one board with a PCI bridge to provide bus compatibility with PCI 2.2 based systems.

It features an optical small form factor (LC) interface that supports shortwave optics and distances up to 500 m at 1 Gb/s and 300 m at 2 Gb/s, depending on the fiber optic cabling used.

5.2.5 LP9002L

The LP9002L 64 bit, 66 MHz Fibre Channel PCI (both low profile and standard short form factor) host adapter provides support for 2 Gb/s Fibre Channel data rates. It also features a Centaur ASIC, 266 MIPS onboard processor and offers 256 KB RAM. The LP9002L features an automatic speed negotiation capability that allows complete compatibility with existing 1 Gb/s Fibre Channel SANs, while allowing seamless upgrades to higher speed 2 Gb/s SANs.

It features an optical small form factor (LC) interface that supports either shortwave or longwave optics and distances up to 10 km at both 1 Gb/s and 2 Gb/s, depending on the type of fibre optic cabling used.

5.2.6 LP9002S

The LP9002S is a 64-bit SBus Fibre Channel HBA which has been optimized for Sun Microsystems SBus servers. It features a Centaur ASIC, 266 MIPS onboard processor and offers 128 KB RAM. The LP9002S features an automatic speed negotiation capability that allows complete compatibility with existing 1 Gb/s Fibre Channel SANs, while allowing seamless upgrades to higher speed 2 Gb/s SANs.

It features an optical small form factor (LC) interface which supports shortwave optics and distances up to 500 m at 1 Gb/s and 300 m at 2 Gb/s depending on the type of fiber optic cabling.

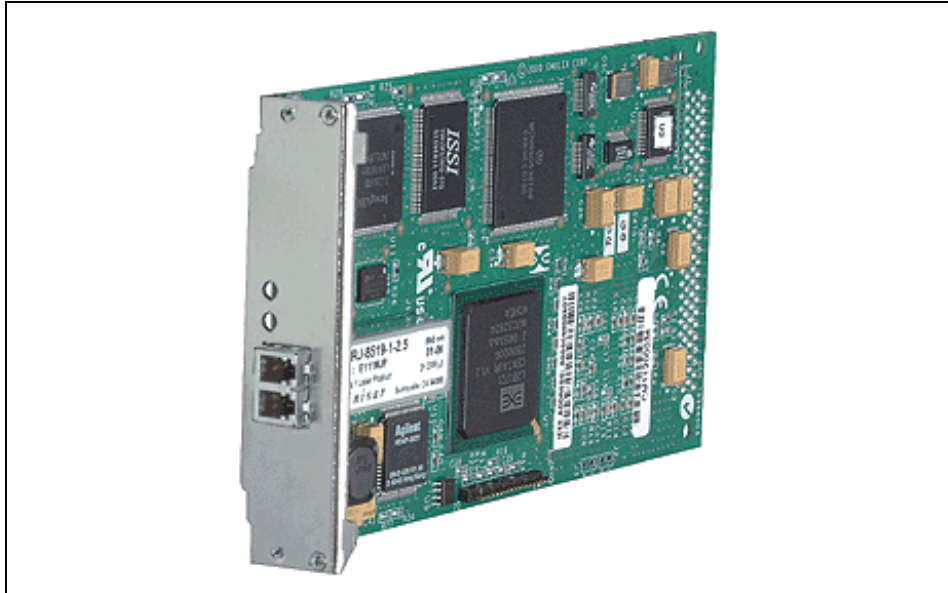


Figure 5-2 Emulex 9002S HBA (SBUS)

For a detailed comparison of all Emulex HBAs, visit Web site:

<http://www.emulex.com/products/white/fc/03-001.pdf>

5.2.7 Emulex special features

In addition to the 64-bit interface, other features unique to the Emulex HBAs are included here.

Persistent binding

This function, available with the Port driver and all UNIX drivers, allows a subset of discovered targets to be bound between a server and device. Binding can be by WWNN or WWPN. Once a configuration has been set, it will survive reboots and hardware configuration changes, as the information will be held in the registry of the server.

For example, this function may be useful for legacy tape software that expects to see its tape devices at the same SCSI Target ID at all times. By binding the tape device's WWN to a SCSI Target ID, we are able to satisfy this criteria.

LUN mapping

This function allows LUNs that are beyond NT's LUN range to be bound permanently to an NT LUN number.

5.2.8 Device drivers

A device driver is a software program that enables a server to communicate with hard drives, CD-ROM drives, printers, and other peripherals. Device drivers are stored on a hard disk and are loaded into memory at boot up.

Emulex provides two device driver options:

- ▶ Fibre Channel port driver
- ▶ Miniport driver

Each of these has a number of utilities.

The device drivers work through a common interface for all Emulex HBAs from the LP7000E through to LP9002S, allowing for a common look and feel across hardware platforms. As the device driver is used to communicate with the server, only one driver version can be loaded onto the server.

Note: IBM supports the Fibre Channel port driver only. Support for the miniport driver can be requested by submitting an RPQ. For this reason we have not included details on the miniport driver.

Fibre Channel port driver

The port driver supports both persistent binding and configurable LUN mapping, and it can map up to 256 LUNs. The driver can also support both FCP and IP on a separate board. The port driver is the only version that allows a floating WWN.

5.2.9 Emulex utilities

Depending on which device driver is selected, there are a number of utilities which enable the setup and modification of the HBA's settings. If the device driver fails to load during boot-up, the utilities will not start. This may occur if a server tried to boot and could not see any attached devices, causing internal conflicts which would prevent the driver from loading.

In this section we detail the function of the port driver utilities.

Elxcfg

Figure 5-3 shows the Emulex Fibre Channel Port Tool (elxcfg.exe) and which is installed automatically as an executable file during the Fibre Channel Port driver installation. When launched, the configuration tool will probe the registry. Adapters defined in the registry are listed in the Available Adapters list box. This list displays the adapter type, bus number, slot location, and firmware revision.

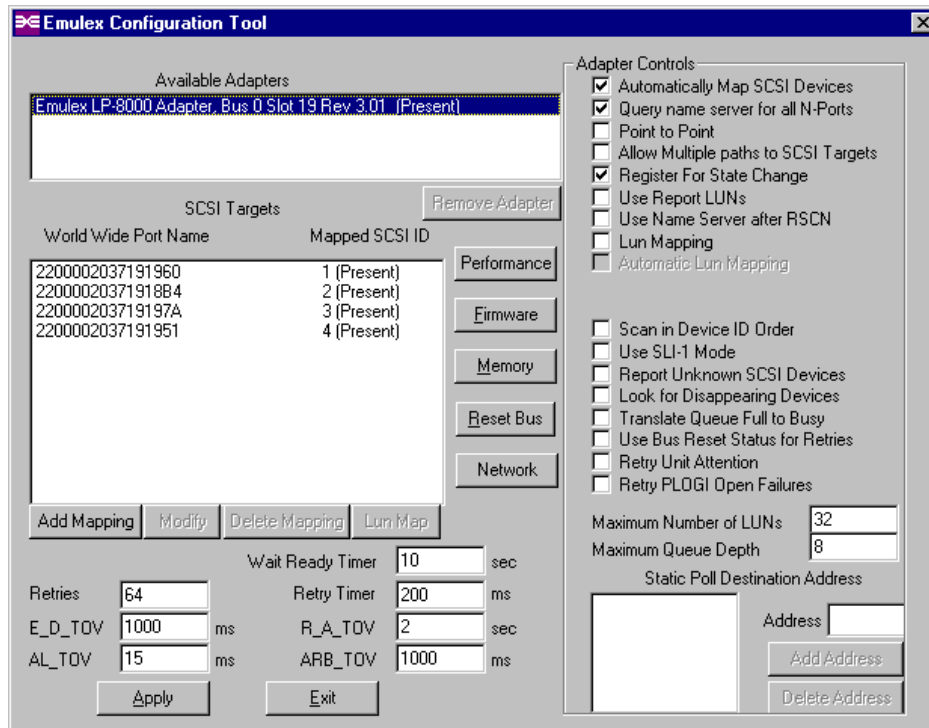


Figure 5-3 Emulex — port driver screen

lp6dutil

lp6dutil will only run in DOS and will not run in a Windows command prompt window. It contains a full suite of diagnostics as well as utility-type functions. It is mainly used for full diagnostics.

This utility is included with the latest firmware download and is common to both driver options.

lputil

lputil is a utility used for UNIX and Novell systems to update host bus adapter firmware and Open Boot code, and lputil is included in driver downloads for each specific UNIX and Novell operating system.

5.2.10 Installation

Every Emulex host adapter is shipped with a unique World Wide Name (WWN), a serial number, and the factory installed firmware level. These numbers are clearly marked on the box and board. We recommend that you record these numbers before installation. Here is a high level overview of the installation process:

- ▶ Power off the server.
- ▶ Insert the HBA into the server.
- ▶ Connect the fabric.
- ▶ Power on the server and run the LED self-test.

5.2.11 Management

Next we discuss two methods of management.

HBAnyware

HBAnyware provides an extensible operating system-independent framework for communication with, and management of, Emulex HBAs. The framework consists of host system resident clients, agents, and services, as well as an Emulex defined command set utilizing the industry-standard Fibre Channel General Service Common Transport (FC-GS-3 CT) protocol as an in-band transport mechanism. Both host-based applications and Fibre Channel attached appliances and switches can utilize the framework. All HBAnyware operations are restricted by an access control list and are cryptographically verified using FC-GS-3 Authenticated CT.

Utilizing HBAnyware, Emulex utilities and third party applications are able to uniformly discover, report on, and manage both local and remote Emulex HBAs from a single management console or application. An integrated set of OS independent APIs, based on OS specific libraries, are included with the framework for use by host-based applications utilizing Emulex driver capabilities. Embedded environments directly attached to the fabric have equivalent capabilities via Emulex's FC-GS-3 CT based command set. In addition, a remote equivalent of the widely used Common HBA API has been added to HBAnyware. With this addition, host-based software management applications will no longer be required to deploy application specific agents to all hosts on the fabric.

MultiPulse

MultiPulse is a driver-based high availability solution that provides failover and dynamic load balancing capabilities when used with Emulex Fibre Channel or iSCSI host bus adapters. MultiPulse supports current multipathing solutions by enabling OEMs to incorporate MultiPulse technology into their own products via Emulex's APIs.

MultiPulse can monitor up to four end-to-end data paths to each configured LUN. Traffic is instantly rerouted around a failed element to protect application availability using all levels of redundancy implemented in the storage network. When a failed path is again made available, the administrator has the option of manually or automatically reinserting the path, which makes it immediately available and restores bandwidth and application availability.

In cluster configurations (such as Microsoft Cluster Server and UNIX-based clusters), MultiPulse prevents resource-induced node failovers, thereby enhancing node and application uptime. In situations where quality of service is imperative, MultiPulse can be configured with hot standby adapters and paths, guaranteeing consistent bandwidth and preventing any service degradation.

5.2.12 Troubleshooting

There are several areas that could be checked to identify a potential problem with the HBA. In the following section we look at some of the areas that should be checked during fault identification and isolation.

Common physical layer problems

Some of the symptoms of other common problems are detailed below:

- ▶ **Bad cable:** There will be a high number of I/O time-outs, the performance may be slow, or the link may be unstable, resulting in a high number of LIPs.
- ▶ **Loose GBIC:** Symptoms similar to a bad cable and may happen when the cable is disconnected and reconnected many times.
- ▶ **Driver set for incorrect topology:** For example, the driver is configured for point-to-point, but the cable is connected to an FL_Port or other FC_AL only device.

If you are unable to resolve a problem and need to refer it to the IBM call center, you should first prepare the following information related to the HBAs:

- ▶ Versions of all Emulex software and firmware
- ▶ LED activity
- ▶ Event logs

Server operating system logs

Error messages are recorded in the standard operating systems log; for Windows this is the event log. For Solaris it is the var/adm/messages log. You can reference the event log and search for any Emulex driver related messages.

The port driver logs events under:

```
elxsl2 event id 1 through 39
```

LED status table

The LED status table, shown in Table 5-1, displays various colors and sequences that can be used to determine if the HBA is functioning correctly.

A properly functioning adapter always has at least one LED flashing. If at least one LED is not flashing, the board is likely hung or dead.

Table 5-1 LED HBA status

Green LED	Yellow LED	State
ON	Slow blink	Link up
	OFF	Link down or adapter not yet configured
OFF	Flickering	Power up or adapter reset
OFF	Fast blink	POST failure
Slow blink	Fast blink	Download in progress or no functional firmware found

A slow blink is 1 per second and a fast blink is 4 blinks per second.

FC port driver

Device driver problems can often be related to an older version of firmware. Always check that you are running the latest level of supported device drivers and firmware.

Refer to the Emulex Web site for the latest version:

<http://www.emulex.com/ts/dds.html>

There are optional driver registry settings that enable extra log information to be collected.

For the FC port driver, the Driver Trace Mask setting could fill the system log, but offers much more flexibility as to what is logged.

There are a number of items that can be individually logged. These values produce a bit map of the item to be logged. For example, if you wanted to trace all of these values:

- ▶ SCSI errors — 0x2
- ▶ Initialization errors — 0x40
- ▶ IOCTL command traffic — 0x80
- ▶ SCSI reset — 0x100
- ▶ Device login trace — 0x200
- ▶ Device response to PRLI — 0x8000

You would add up these values and place the result in the Driver Trace Mask value:

Driver Trace Mask = 0x83b2

Note: Some of these values (0x1, 0x8) can potentially trace every command to and from the driver, which can potentially fill the log.

5.2.13 Performance

Some of the settings that affect performance are detailed in the following topics.

Driver settings

There are a number of adapter timer settings that can be modified. IBM recommends leaving all settings to the defaults provided by Emulex. In Figure 5-4 we show the adapter panel settings.

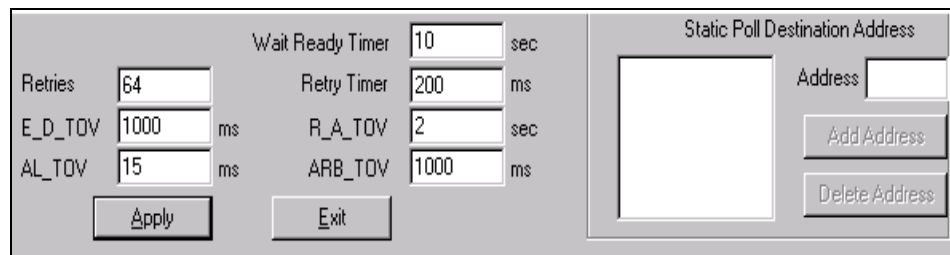
The image shows a screenshot of the 'Emulex adapter settings' window. It contains several input fields for configuration. On the left, there are fields for 'Retries' (64), 'E_D_TOV' (1000 ms), and 'AL_TOV' (15 ms). In the center, there are fields for 'Wait Ready Timer' (10 sec), 'Retry Timer' (200 ms), 'R_A_TOV' (2 sec), and 'ARB_TOV' (1000 ms). At the bottom left are 'Apply' and 'Exit' buttons. On the right, there is a section titled 'Static Poll Destination Address' which includes a large empty box for the address, an 'Address' input field, and 'Add Address' and 'Delete Address' buttons.

Figure 5-4 Emulex adapter settings

An exception to default settings would be when a server is connected to multiple storage devices. Some hardware vendors have a requirement to alter the default value settings, and they must be consulted.

For example, the recommended R_A_TOV value for one vendor is double that of the Emulex default.

R_A_TOV and E_D_TOV

These two parameters change the Resource Allocation Time-out Value and Error Detect Time-out Value respectively. If a switch is present, these values will be obtained from the switch, thus overriding any values entered in the configuration tool or registry.

HBA settings

HBA settings vary from model to model and might be set differently, depending on the platform and operating system. In Table 5-2 and Table 5-3 we show some samples of HBA settings that can be set. For current settings and specific recommendations, refer to the IBM SSG HBA and SAN Interoperability Matrix at the Web site:

http://ssddom02.storage.ibm.com/hba/hba_support.pdf

Table 5-2 LP8000 adapter recommended configuration file parameters

Parameters	Recommended settings
automap	2: Default. Automatically assigns SCSI IDs to Fibre Channel protocol (FCP) targets.
fcv-on	1: Default. Turn on FCP.
lun-queue-depth	16: Recommended when there are less than 17 LUNs per adapter. Set value = $256 \div (\text{total LUNs per adapter})$ when there are more than 16 LUNs per adapter. If your configuration includes more than one LP8000 adapter per server, calculate the LUN-queue-depth value using the adapter with the most LUNs attached.
no-device-delay	1: Recommended. Delay to failback and I/O.
network-on	0: Default. Recommended for fabric. Do not turn on IP networking. 1: Turn on IP networking.
scan-down	2: Recommended. Use an inverted ALPA map and create a target assignment in a private loop.
topology	2: Recommended for fabric. Point-to-point topology only. 4: Recommended for nonfabric. Arbitrated-loop topology only.
zone-rscn	0: Default 1: Recommended for fabric. Check name server for RSCNs.

Table 5-3 LP9000 adapter recommended configuration file parameters

Parameters	Recommended settings
automap	1: Default. SCSI IDs for all FCP nodes without persistent bindings will be automatically generated. If new FCP devices are added to the network when the system is down, # there is no guarantee that these SCSI IDs will remain the same # when the system is booted again. If one of the above fcp binding methods is specified, then automap devices will use the same mapping method to preserve CSI IDs between link down and link up. If no bindings are specified above, a value of 1 will force WWNN binding, 2 for WWPN binding, and 3 for DID binding. If automap is 0, only devices with persistent bindings will be recognized by the system.
fcp-on	1: Default. Turn on FCP.
lun-queue-depth	30: The default value lpfs will use to limit the number of outstanding commands per FCP LUN. This value is global, affecting each LUN recognized by the driver, but may be overridden on a per-LUN basis. RAID may want to be configured using the per-LUN tunable throttles.
no-device-delay	0: Default. Implies no delay whatsoever. 1: Recommended. 2: Setting a long delay value may permit I/O to build up, each with a pending timeout, which could result in the exhaustion of critical Solaris kernel resources. In this case, you may see a fatal message such as PANIC: Timeout table overflow
network-on	0: Default. Recommended for fabric. Do not turn on IP networking. 1: Turn on IP networking.
scan-down	0: Recommended. Causes the lpfs driver to use an inverted ALPA map, effectively scanning ALPAs from high to low as specified in the FC-AL annex. 2: Arbitrated loop topology.
tgt-queue-depth	0: Recommended. The default value lpfs will use to limit the number of outstanding commands per FCP target. This value is global, affecting each target recognized by the driver, but may be overridden on a per-target basis (see below). RAID may want to be configured using the per-target tunable throttles.
topology	2: Recommended for fabric. Point-to-point topology only. 4: Recommended for nonfabric. Arbitrated-loop topology only.
xmt-que-size	256: Default. Size of the transmit queue for mbufs (128 - 10240).
zone-rscn	0: Default 1: Recommended for fabric. Check name server for RSCNs. Setting zone-rscn to 1 causes the driver to check with the NameServer to see if an N_Port ID received from an RSCN applies. If soft zoning is used with Brocade fabrics, this should be set to 1.

External boot function

Emulex HBAs provide the ability to perform an external server boot from a Fibre Channel device. The boot BIOS is disabled by default and must be enabled with the Ip6dutil or GUI utility. You can define up to eight boot devices configured by Device ID or WWPN. The boot BIOS supports up to eight FC HBAs per server.

Firmware structure

Emulex provides two versions of zipped firmware; an AWC and DWC. The AWC file will update all layers of the firmware structure except for the Config layer. The DWC updates the same layers as the AWC with the exception of the Adapter Boot - POST layer. The Adapter BOOT- POST contains a kernel and is essential for normal operation.

As an AWC firmware load will update the Adapter BOOT - POST region, if the firmware load is interrupted during an update to the kernel, it could destroy the HBA. For this reason it is not recommended to download an AWC firmware load directly to the HBA. With the DWC firmware load, if the update process is interrupted, the POST code is not affected and the download process can be retried. The firmware readme file contains information on the regions and should be referred to prior to the load.

Typically, major firmware changes, for example from 1.0 to 2.0, will require an AWC load. For interim changes, for example from 2.0 to 2.1, refer to the readme file to determine if an AWC or DWC load is required.

Note: For all other firmware loads, it is recommended to perform a DWC load.

In Table 5-4 we describe each layer of the firmware.

Table 5-4 *Emulex firmware structure*

Layer	Description
Adapter Boot - POST	Contains kernel — essential for normal operation
ENDEC Loop Back	POST code for the internal ENDEC loop back
Stub	Loads either SLI-1 or SLI-2 function firmware
Boot BIOS (optional)	Optional INT13 boot BIOS
SLI - 1 Overlay	SLI-1 functional firmware
SLI - 2 Overlay	SLI-2 functional firmware
Config Regions	Non-volatile configuration parameters

5.3 JNI

IBM currently also markets a broad range of Fibre Channel HBAs from JNI. These HBAs operate in a number of operating systems environments, including Solaris, Microsoft Windows in multiple server and cluster environments, Hewlett Packard HP-UX, IBM AIX, Red Hat Linux, Novell NetWare, and Apple Mac OS. In the following sections we have listed the IBM supported offerings.

5.3.1 FCI-1063-N 32-bit PCI to FC HBA

The JNI FCI-1063 is a 32-bit PCI-to-FC Adapter with an Integrated Optical Short-Wave Dual SC Connector Interface. The FCI-1063-N provides a full-duplex 1.0623 Gb Fibre Channel connection between PCI Sun servers and SAN devices.

5.3.2 FC64-1063-N 64-bit SBus to FC HBA

The JNI FC64-1063-N 64-bit SBus-to-FC Adapter with an Integrated Optical Short-wave Dual SC Connector Interface. The FC64-1063-N provides a full-duplex 1.0623 Gb Fibre Channel connection between SBus Sun servers and SAN devices. In Figure 5-5 we show a picture of the FC64-1063-N HBA.

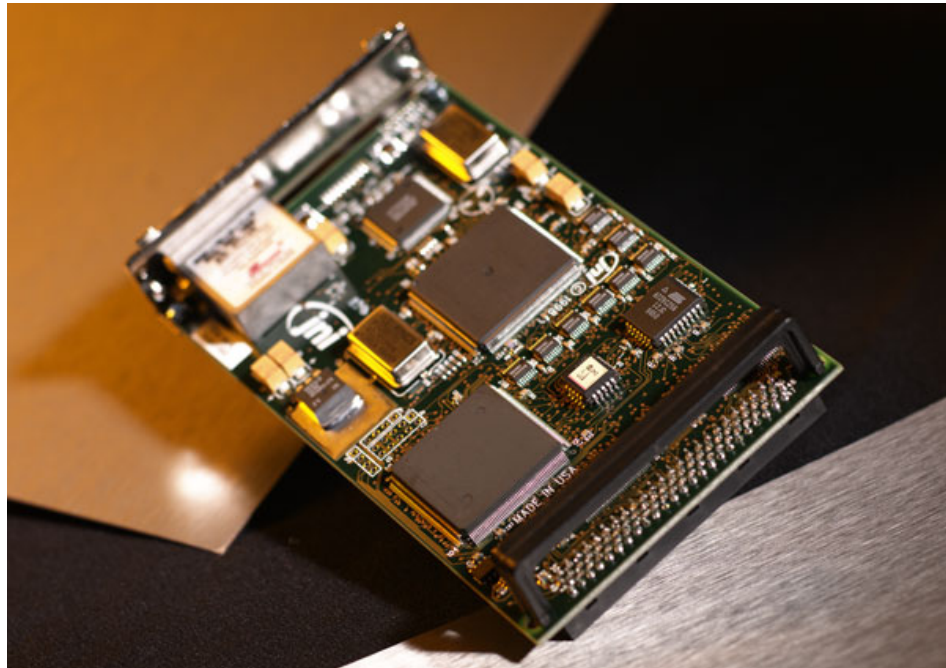


Figure 5-5 Picture of JNI's FC64_1063 HBA

5.3.3 FCE-1063 64-bit SBus to FC HBA

The JNI FCE-1063 is a 64-bit SBus-to-FC Adapter with an Integrated Optical Short wave SC Connector Interface. The FCE-1063 is backward-compatible with 32-bit SBus Sun servers and provides a full-duplex 1.0623 Gb Fibre Channel connection between SBus Sun servers and SAN devices.

5.3.4 FCE2-1063 64-bit Dual Port SBus to FC HBA

The JNI FCE2-1063 is a 64-bit Dual Port SBus-to-FC Adapter with Integrated Optical Short wave Dual SC Connector Interfaces. The FCE-1063 supports 32-bit or 64-bit data transfer paths and provides a full-duplex 1.0623 Gb Fibre Channel connection between SBus Sun servers and SAN devices.

5.3.5 FCE-1473 1 or 2 Gb 64-bit SBus to FC HBA

The JNI FCE-1473 is a 64-bit Single Port SBus-to-FC Adapter with an Integrated Short wave LC small factor Optical Interface. The FCE-1473 supports 32-bit and 25 MHz and 64-bit data paths. The FCE-1473 is a full featured 1 Gb/s or 2 Gb/s full duplex HBA that supports automatic rate negotiation seamlessly on demand. The FCE-1473 also features both local and fabric boot capabilities.

5.3.6 FCE-6410 64-bit PCI to FC HBA

The JNI FCE-6410 is a 64-bit PCI-to-FC Adapter with an Integrated Optical Short-wave SC Connector Interface. The FCE-6410 supports 32-bit and 64-bit data paths at 33 MHz and provides a full-duplex 1.0623 Gb/s Fibre Channel connection between multiple OS (IBM AIX, Red Hat Linux, Novell Netware, Sun Solaris, HP-UX, Windows, Mac OS) servers, and SAN devices. The FCE-6410 provides multiple OS support through two JNI products (UNIX DriverSuite, PC DriverSuite), bundled with the HBA.

5.3.7 FCE2-6412 64-bit Dual Port PCI to FC HBA

The JNI FCE2-6412 is a 64-bit Dual Port PCI-to-FC Adapter with Integrated Optical Short-wave Dual SC Connector Interfaces. The FCE-6412 supports 64-bit data paths at 33 MHz and 66 MHz and provides a full-duplex 1.0623 Gb/s Fibre Channel connection between multiple OS (IBM AIX, Red Hat Linux, Novell Netware, Sun Solaris, HP-UX, Windows, Mac OS) servers, and SAN devices. The FCE2-6412 provides multiple OS support through two JNI products (UNIX DriverSuite, PC DriverSuite), bundled with the HBA.

5.3.8 FCC-6460 1 or 2 Gb Compact PCI to FC HBA

The JNI FCC-6460 is a 64-bit Dual Port PCI-to-FC Adapter with an Integrated Optical Short-wave LC small form factor Connector Interface. The FCC-6460 supports 64-bit data paths and is a full featured 1 Gb/s or 2 Gb/s full duplex HBA between Sun Servers and SAN devices and supports automatic rate negotiation seamlessly on demand. The FCC-6460 is also hot swappable.

5.3.9 FCE-6460 1 or 2 Gb 64-bit PCI to FC HBA

The JNI FCE-6460 is a 64-bit Single Port PCI-to-FC Adapter with an Integrated Optical Short-wave LC small form factor Connector Interface. The FCE-6460 supports 64-bit data paths at 33 MHz or 66 MHz and is a full featured 1 Gb/s or 2 Gb/s full duplex HBA that supports automatic rate negotiation seamlessly on demand. The FCE-6460 also contains SNIA library support for management applications in Solaris or Windows.

Note: All JNI Fibre Channel HBAs support fabric boot capabilities.

5.3.10 Drivers

Both the FCI_1063 and FC64_1063 run the Solaris driver. This driver supports 2.6, 7, and 8 versions of SUN Solaris. **fca-pci.pkg** is the SCSI driver package used in Solaris with the JNI 32-Bit PCI adapter and **fcaw.pkg** is used with the 64-bit SBus adapter.

To view the adapter's properties, type in a **show-devs** command to determine the location of your adapter (remember you are looking for 1242, 4643 or Fibre Channel). Now **cd** to the directory where the JNI adapter card is located. Once you **cd** to the directory, type in the command **properties**. Make sure you set the Open Boot parameter auto-boot to false. Remember that you must be connected to a device or have a loopback plugged into the adapter before you can view the adapter properties.

Prior to installing the driver, you can enter the following command to perform a basic test to determine if the Open Boot can recognize the JNI adapter card:

Show-devs

LUN level masking

With the JNI EZ Fibre software you have the ability to implement LUN-Level Zoning at the host bus adapter level. The LUN-Level Zoning option comes standard on all JNI adapter cards and offers the following advantages:

- ▶ Shorter boot-up time by controlling device discovery process in multiple CPU environments
- ▶ Flexibility (dynamic allocation, the ability to change drives “on the fly”)
- ▶ Allocation of backup resources (one can allocate backup within a tape array itself)
- ▶ Zero % performance loss (since the zoning has been pre-configured, the operating system does not incur the overhead of determining resource availability)
- ▶ Enhanced security using JNI’s host-based LUN-Level Zoning

5.3.11 Management

JNI’s proprietary software, EZ Fibre, is used for configuring and managing a Fibre Channel installation. EZ Fibre is a Windows-based program that comes with on-screen help, a troubleshooting guide, and customer support options. In Figure 5-6 we show the EZ Fibre configuration utility, which is the fastest and easiest way to install and configure JNI host bus adapters.

EZ Fibre allows you to link your servers to RAIDs, JBODs, and other storage devices through a high-speed Fibre Channel network. You can map the SAN fabric all the way down to the LUN level, while viewing a graphical interface of the Fibre Channel devices attached to JNI adapters.

EZ Fibre dynamically discovers new targets attached to the FC link, and through its GUI allows the administrator to easily set and manage the host bus adapter parameters. Such “hot-pluggable” capabilities minimize the steps necessary to grow a SAN or swap out a bad disk drive.

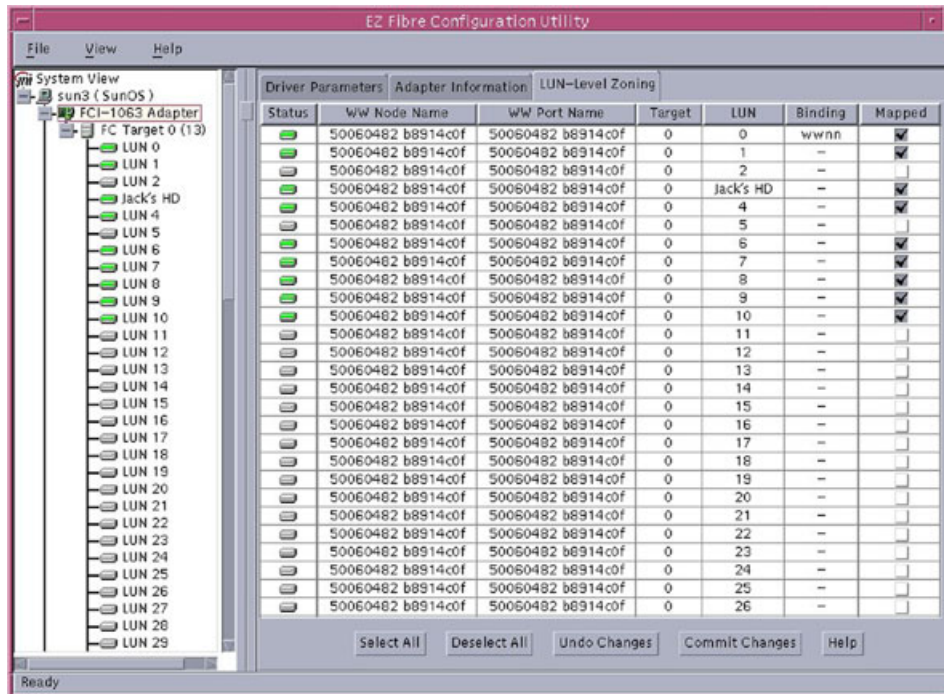


Figure 5-6 EZ Fibre configuration panel

5.3.12 Installation

Every JNI host bus adapter is shipped with a unique World Wide Name (WWN). These numbers are currently located via the EZ Fibre software shipped with the cards.

We recommend that you record these numbers before installation.

Before installing the adapter card, ensure that the adapter is of the correct type. The adapter will only operate when connected to devices of the same type. In Figure 5-7 we show the FCI-1063-N.

The standard SBUS back plate includes a two-tab extender bar. Check your system to see if the adapter bar is required. If the bar needs to be removed, use the screwdriver supplied to remove the two small Phillips head screws, which mount the bar to the back plate.

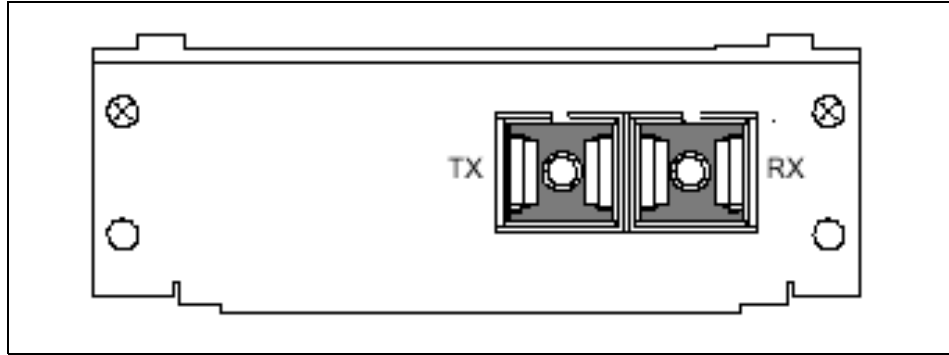


Figure 5-7 The FCI-1063_N 64 Optical device

Next we describe is a high-level overview of the installation process. There are slight differences, depending on which model of SUN server you are configuring:

- ▶ Power off the server
- ▶ Insert the HBA into the server
- ▶ Connect fabric
- ▶ Power on the server

5.3.13 Troubleshooting the SBUS HBA

In this section, we show some troubleshooting hints and tips for the SBUS HBA.

Is dynamic reconfiguration supported?

The fcaw driver supports Dynamic Reconfiguration.

I cannot see the adapter with show-sbus or show-devs.

Verify that the adapter is properly seated, then turn the power on. Check the probe-list by typing in the command **printenv** and look for the Open Boot parameters **sbus-probe-list** or **sbus-probe-default**.

Two columns should appear. The first column displays the slots on the computer that are being probed. The second column displays all possible slots to be probed. Compare the two columns and find the slot that is not being probed. Enter slot into the probe list. To reset the probe list to include the slots, use the command:

```
setenv sbus-probe-list or sbus-probe-default = (NNNN)
```

“NNNN” represents all slots to be probed.

5.3.14 Troubleshooting the JNI PCI HBA

In this section, we give some troubleshooting hints and tips for the PCI HBA.

I cannot see the adapter with show-sbus or show-devs command.

Verify that the adapter is properly seated, then turn power on. Check the probe-list by typing in the command **printenv** and look for the Open Boot parameters **pci-probe-list**.

Two columns should appear. The first column displays the slots on the computer that are being probed. The second column displays all possible slots to be probed. Compare the two columns and locate the slot that is not being probed. Enter the slot into the probe list. To reset the probe list to include the slot, use the command:

```
setenv pci-probe-list = (NNNN)
```

“NNNN” represents all slots to be probed. Make sure the PCI adapter is in the correct slot. It must be 33MHz, 5 volt and 2.1 PCI compliant.

I get an error that the driver fails to initialize.

Check **fca-pci.conf** for **fca_nport**. In Solaris view the **fca-pci.conf**. The file is found in the **/kernel/drv** directory-ry. Type in **more fca-pci.conf**. Verify that the parameter **fca_nport** is correctly set. Remember that the configuration flag **fca_nport** is a Boolean Type value. The default is 0 (false). If false (0), fca initializes on a loop. If true (1), fca initializes as an N_Port and fabric operation is enabled. The parameter looks like this:

```
fca_nport = 0;
```

Also check that the drives are attached correctly, or if the cables are faulty. You may also receive continuous Elastic Store errors from the driver.

5.3.15 Troubleshooting both PCI and SBUS

In this section we look at some common problems and detail some specific settings IBM recommends when connecting to the ESS.

The drives were not added.

Perform the **drvconfig** and **disks** command.

Check the file **/kernel/drv/sd.conf**. This file determines the targets and LUNs the target drive will probe for. If the targets/LUNs you are attaching do not appear, edit the file and add them. Halt the system and perform a **boot-r**.

I do not see all my targets and LUNs.

You must have all targets and LUNs in the **sd.conf** file for Solaris to recognize them. The **sd.conf** file is found in the **/kernel/drv**. By default you will have targets 0-6 and 7-15 with one LUN under each target.

Here is an example of how each line should look:

```
name=l.sd1T class=lsscsi1@ target=N lun=0; (where N is the target number)
```

Enter this number into **sd.conf** with the appropriate number of LUNs. There should be one LUN under each target. A reconfiguration boot is required after editing the **sd.conf** file before the device is recognized by Solaris. Type in the Solaris command **drv- config**. This command goes out and looks for new devices attached to the JNI adapter cards (remember that the targets and LUNs must be configured in the **sd.conf** before trying this command). After the **drv-config** command, you will receive a new prompt. At this prompt, type in the command **disks** and press Enter. Verify the presence of the new targets by entering the **format** command.

Disk problems can also affect device recognition.

I cannot see my targets over a switch.

Make sure the cables are correctly plugged in and there is a link (on the switch). Make sure that in the configuration file (**fcaw.conf**), the parameter **fca_nport** is set to 1.

As an example:

```
Configuration flag fca_nport
Type: Boolean; default: 0 (false)
If false (0), then fca initializes on a loop
If true (1), then fca initializes as an N_Port and fabric operation is enabled
fca_nport = 1;
```

After I reboot, I receive Target_Queue full error.

The **sd_max_throttle** variable is the maximum number of commands that the SCSI sd driver will attempt to queue to the HBA driver (fcaw). The default value is 256. If **sd_max_throttle** is set at its default you will receive the error:

```
PCI - fca-pci0: fca_highintr: Target Queue Full. Packet Rejected!
SBUS - fca0: fca_highintr: Target Queue Full. Packet Rejected!
```

Refer to “sd_max_throttle” on page 172 for the IBM recommended setting, and download the JNI driver 2.4 or later. This driver fixes the queue-full condition.

5.3.16 Performance settings

In this section we look at the recommended IBM values for the FCI-1063 and FC64-1063 when connecting to an ESS. These values can be located in the manual *IBM TotalStorage ESS 2105*, SC26-7446.

System settings

All values can be located in the Systems file, which can be found in the `/etc` directory.

The two parameters that need to be edited are the `sd_max_throttle` and the `maxphys`.

`sd_max_throttle`

This parameter specifies the maximum number of commands that the sd driver will queue to the host bus adapter driver. The recommended value is 16 for Fibre Channel configurations with less than 17 LUNs per adapter. The default value is 256, but you must set the parameter to a value less than or equal to the maximum queue depth for each LUN connected. For configurations with greater than 16 LUNs per adapter, use the following formula for each adapter:

```
sd_max_throttle=256 / (LUNs per adapter)
```

Where LUNs per adapter is the largest number of LUNs assigned to a single adapter.

Use the adapter with the highest number LUNs attached when calculating the lun-queue-depth value for servers with more than one adapter installed.

```
set sd:sd_max_throttle=5
```

If you are using version 2.4 of the driver or later, **`sd_max_throttle`** need not be reduced. The target driver (fcaw) will respond by single-threading all subsequent I/O, which has a negative impact on performance (when Solaris overloads the target with commands, the target will reject subsequent SCSI commands because its internal queues will register as full. The Solaris response will be to only send one command at a time — which in turn will lower performance).

`sd_io_time`

This parameter specifies the time-out value for disk operations. Add the following lines to the `/etc/system` file to set the `sd_io_time` parameter for the ESS LUNs:

```
set sd:sd_io_time=0x78
```

sd_retry_count

This parameter specifies the retry count for disk operations. Add the following lines to the **/etc/system** file to set the **sd_retry_count** parameter for the ESS LUNs:

```
set sd:sd_retry_count=5
```

maxphys

This parameter specifies the maximum number of bytes you can transfer for each SCSI transaction. The default value is 126976 (124 KB). If the I/O block size requested exceeds the default value, the request is broken into more than one request. The value should be tuned to the intended use and application requirements. For maximum bandwidth, set the **maxphys** parameter by adding the following line to the **/etc/system** file:

```
set maxphys=8388608
```

If you are using Veritas volume manager on the ESS LUNs, you must set the VxVM max I/O size parameter, (**vol_maxio**) to match the **maxphys** parameter. For example, if you set the **maxphys** parameter to 8388608 you will need to add the following line to the **/etc/system** file to set the **VxVM I/O** size also to 8 MB:

```
set vxio:vol_maxio=16384
```

HBA settings

HBA settings vary from model to model and might be set differently depending on the platform and operating systems. These are samples of HBA settings that can be set. For current settings and specific recommendations refer to the IBM SSG HBA and SAN Interoperability Matrix at the Web site:

http://ssddom02.storage.ibm.com/hba/hba_support.pdf

Table 5-5 IBM recommended settings for JNI FC64-1063 and JNI FCI-1063

Parameters	Recommended settings
FcLoopEnabled FcFabricEnabled	For direct attachment, set FcLoopEnabled=1 and FcFabricEnabled=0, For fabric attachment, set FcLoopEnabled=0 and FcFabricEnabled=1
fca_nport	0 = default, initializes on a loop. 1 = recommended for fabric, initializes as an N-Port
public loop	0 = default, recommended, initialize according to what fca_nport is set for disabled
ip_disable	0 = default, IP side of the driver is enabled. 1 = recommended for fabric, IP side of the adapters is completely disabled.

Parameters	Recommended settings
failover	60 -recommend without McDATA switch, 300 -recommended with McDATA switch.
busy_retry_delay	500 -recommended, delay between retries after device returns busy response for a command.
scsi_Probe_delay	5000 -recommended, delay before SCSI probes are allowed during boot

In Table 5-6 we show settings for other JNI adapters.

Table 5-6 JNI miscellaneous recommended settings

Parameters	Recommended settings
FcEngHeartbeatInterval	5: Default. When the JNI adapter/driver detects that the Fibre Channel link is up (and there is no I/O activity), it will send a test frame (or heartbeat) to itself to verify link integrity. The test frame is sent at the interval specified by this parameter. If the test frame does not complete, it is assumed that there is a link problem. In this situation, the driver initiates error recovery to re-establish a good link. A value of 0 disables the heartbeat
FcLinkUpRecoveryTime	1000: Default. Delay (msec) after the link is up before port discovery begins, allowing the link to stabilize and protecting against a possible I/O surge. This timer is reset every time the link comes up. The default value is adequate for most configurations.
BusyRetryDelay	5000: Default. Delay (msec) before retrying after receipt of an I/O with a SCSI Busy status from a target. The number of retries is based on the Solaris retry count associated with the I/O.
FailoverDelay	30: Delay (seconds) before failing all I/O for an offline target. If the delay timer expires, all I/O for the failed target is returned to the application. A zero value disables failover.
TimeoutResetEnable	0: False. Boolean parameter for enabling SCSI target resets for timed out I/O. When the timer expires (usually 60 seconds, as specified by the upper layers), the driver issues a target reset to attempt to clear the device (which might be either too busy to respond or stuck).

Parameters	Recommended settings
QfullRetryCount	5: Default. Number of times an I/O is retried due to receipt of a SCSI queue full status from a target. The delay between retries is based on the QfullRetryDelay parameter.
QfullRetryDelay	5000: Default. Delay (msec) before retrying after receipt of an I/O with a SCSI queue full status from a target. The number of retries is based on the QfullRetryCount parameter.
LunRecoveryInterval	50: Default. Sets the LUN I/O recovery interval (in msec) after the driver reconnects to a disk. It is a global parameter affecting all targets, and determines how long the driver waits after a port is discovered until sending I/O to that port. Some devices might require more time to flush I/O that was in progress prior to a link going down; if this is the case, increase the value of this parameter.
FcLinkSpeed	3: Default. Specifies the desired Fibre Channel link speed as follows: v 0: default to SEEPROM setting v 1: force 1 gigabit per second v 2: force 2 gigabit per second v 3: auto negotiate link speed
JniCreationDelay	5: Default. Delay (seconds) after driver creation to allow the network to stabilize, discover ports, and build the driver's database. Increase this value if targets are being discovered too late in the boot process.
FlogiRetryCount	3: Default. Total number of Fabric Login (FLOGI) attempts before giving up logging in to a switch. Failure prevents participation on a Fabric topology.
FcFlogiTimeout	10: Default. Specifies the amount of time (in seconds) that the driver waits for a Fabric Login (FLOGI) accept. The value should be increased only if the switch to which the HBA is connected requires more time to respond to a FLOGI. The number of retries is configured with the FlogiRetryCount parameter.

Parameters	Recommended settings
PlogiRetryCount	5: Default. Total number of Port Login (PLOGI) attempts before giving up logging in to a SCSI target.
PlogiControlSeconds	30: Default. Defines the number of seconds that the driver waits for a successful port login (PLOGI) attempt. The maximum number of attempts is defined by the PlogiRetryCount parameter. Some devices might take longer to respond to PLOGIs; if this is the case, increase the value of this parameter.
FcEmldEngTcbCount.	1789: Default. Total number of concurrent exchanges (also called transfer control blocks) allowed by the adapter. To optimize performance, set this parameter to match the memory capacity of the hardware

Boot BIOS

All JNI HBAs support external boot.

5.4 QLogic

IBM provides support for the QLogic family of Fibre Channel HBAs. This includes specific HBAs in the QLA2100, QLA2200 and QLA 2300 series. QLogic Fibre Channel HBA products have achieved SANMark certification, which is the industry standard for device compatibility. The HBAs are based on single chip architecture, providing high reliability and low power consumption. In addition, they are able to boot to an external FC storage device, either on a local loop or through a switched fabric.

For a complete list of supported drivers and configurations, refer to the Web site:

<http://www.qlogic.com>

5.4.1 QLA2100

QLogic's first generation of Fibre Channel HBA products is the QLA2100 family. These cards are currently available as either fixed copper (QLA2100/66) or optical (QLA2100F/66) nodes, IBM only supports the optical version. The /66 indicates the maximum supported PCI bus speed. These are 64-bit PCI cards that also function in 32-bit PCI environments. The HBAs use the ISP2100 ASIC.

The 2100 HBAs operate at 1Gb/s data rate on the Fibre Channel medium. These products support either FC-AL or switched fabrics using an FL_Port connection.

In Figure 5-8 we show the QLogic HBA card.

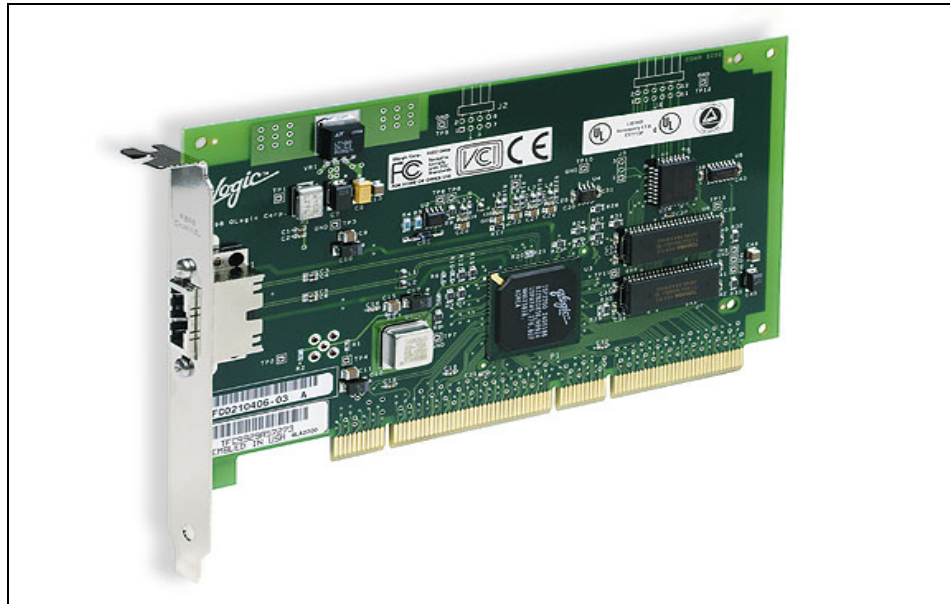


Figure 5-8 QLogic HBA card

5.4.2 QLA2200

QLogic's second generation of performance optimized Fibre Channel HBA's is the QLA2200 series. These boards are based on QLogic's ISP2200 ASIC. As with the 2100 series, these are PCI cards that operate in 33 and 66 MHz as well as 32 and 62-bit environments. The 2200 series supports FC-AL as well as switched fabric via F_Port and FL_Port connections. Additionally, the QLA2200 series is able to support IP protocol. The QLA2200 series supports Class 2, 3, and FC Tape. This family of boards is available in the following configurations: IBM supports the QLA2200F, QLA2202F for connections to the ESS.

Note: QLA2202 is supported only on the NAS 300G R2.5.

For a detailed description of each QLA2200 HBA, refer to the following Web site:

http://www.qlogic.com/products/sanblade/sanblade_2200.asp

5.4.3 QLA2300

The QLogic SANblade 2300 Series Fibre Channel host bus adapters (HBAs) offer 2 Gb/s performance. They are available in PCI-X form factor, which is backwardly compatible with PCI. SANblade 2300 Series HBAs have an integrated RISC processor, the fibre protocol engine and transceivers in a single, Fibre Channel controller chip. Each HBA features 256 KB RAM per port, supports FC-AL, FCAL-2, point-to-point, switched fabric and Class 3 service. Supported HBAs include the QLA2300F, QLA2310FL and the QLA2340L.

Note: QLA2300F is supported for a Brocade fabric only.

For a detailed description of each QLA2300 HBA, refer to the following Web site:

http://www.qlogic.com/products/sanblade/sanblade_2300.asp

5.4.4 Installation

Refer to the *IBM Enterprise Storage Server Host Systems Attachment Guide, 2105 Models E10, E20, F10 and F20, SC26-7296* for specific installation instructions.

5.4.5 Management

QLogic offers management software with every HBA. IBM also offers a customized version called FasT-MSJ. The QLogic SANsurfer Tool Kit CD is included with every HBA. Contained on the CD is SANsurfer software that simplifies management by making setup, configuration and maintenance easy along with enabling features such as remote management, load balancing, failover and persistent binding. In addition, the easy to use SANblade Control FX wizard based utility is provided for installation, configuration and diagnostic functions. SANblade Control FX is delivered with your drivers. The SANblade 22xx/23xx is also fully compatible with management applications that support the SNIA API, allowing IT managers to reduce HBA management time and lower their total cost of ownership.

5.4.6 Troubleshooting

For detailed debugging, experienced users may prefer to use the Event Viewer that ships with Windows.

Events logged by the driver are listed with the Source field set to “ql2100” or “ql2200” and the Event field set to “11”. Double-clicking on the event entry will

allow you to view the event details, then set the data format to “Words”. The detailed event code is displayed at hexadecimal offset x’34’.

For some of the event codes, additional data will be recorded in the least significant 16 bits of the long word. Additional data may also be recorded in the long word at offset 10 (hex).

There is a FAST!UTIL parameter to enable “additional event logging”, the default is disabled.

Events that are logged without extended event logging being enabled are “true” error events (those without a “*” or “***” in front of them). In general, these error events are logged because of some error conditions. Contact QLogic customer support if you encounter these error events.

5.4.7 Performance

All QLogic PCI cards store user configurable data in the hardware. HBA settings vary from model to model and might be set differently depending on the operating system. These are samples of HBA settings that can be set. For current settings and specific recommendations refer to the IBM SSG HBA & SAN Interoperability Matrix:

http://ssddom02.storage.ibm.com/hba/hba_support.pdf

These may include the following.

Frame Size

This setting specifies the maximum frame length, the default sized for the LA22xx board is 1024 and 2048 for the LA23xx board. If using F_Port (point-to-point) connections, change this setting to 2048 for maximum performance.

Loop Reset Delay

After resetting the loop, the firmware refrains from initiating any loop activity for the number of seconds specified in this setting. The default is 5 seconds.

Adapter Hard Loop ID

This setting forces the adapter to attempt to use the ID specified in the Hard Loop ID setting. The default is disabled.

Hard Loop ID

If the Adapter Hard Loop ID setting is enabled, the adapter attempts to use the ID specified in this setting. The default ID is 0 (disabled).

Execution Throttle

This setting specifies the maximum number of commands executing on any one port. When a port's execution throttle is reached, no new commands are executed until the current command finishes executing. The valid options for this setting are 1-256. The default (optimum) is 16.

LUNs per Target

This setting specifies the number of LUNs per target. Multiple LUN support is typically for RAID boxes that use LUNs to map drives. Options include 0,8,16,32,64,128 or 256. The default is 8. If you do not need multiple LUN support, set the number of LUNs to 0.

Enable LIP Reset

This setting determines the type of loop initialization process (LIP) reset that is used when the operating system initiates a bus reset routine. When this setting is Yes, the driver initiates a global LIP reset to clear the target device reservations. When this setting is No, the driver initiates a global LIP reset with full login. The default is No.

Enable LIP Full Login

This setting instructs the ISP chip to re-login to all ports after any LIP. The default is enabled.

Enable Target Reset

This setting enables the drivers to issue a Target Reset command to all devices on the loop when a SCSI Bus Reset command is issued. The default is disabled.

Login Retry Count

This setting specifies the number of times the software tries to log in to a device. Options are 0-255, the default is 8 retries.

Port Down Retry Count

This setting specifies the number of times the software retries a command to a port returning port down status. Options are 0-255, the default is 8 retries.

Extended Error Logging

This setting provides additional error and debug information to the operating system. When enabled, events are logged into the Windows NT Event Viewer. The default is disabled.

Operation Mode

This setting specifies the reduced interrupt operation (RIO) modes, if supported by the software driver. RIO modes allow posting multiple command completions

in a single interrupt (see the QLogic hardware install manual for complete description of usage). The manuals are available on the QLogic Web site. The default is 0 disabled.

Connection Options

This setting defines the type of connection (loop or point-to-point) or connection preference: 0 = loop only, 1 = point-to-point only, 2 = loop preferred, otherwise point-to-point, 3 = point-to-point, otherwise loop. The QLogic default is 2.

Interrupt Delay Timer

This setting contains the value (in 100-microsecond increments) used by a timer to set the wait time between accessing (DMA) a set of handles and generating an interrupt. The default is 0. Only applies the QLA23xx adapter.

In Table 5-7 we show the recommendations for the QLA2200F.

Table 5-7 QLA2200F recommendations

Parameters	Recommended settings
hba0-max-frame-length	2048
hba0-max-iocb-allocation	256
hba0-execution-throttle	31
hba0-login-timeout	4
hba0-login-retry-count	1
hba0-fabric-retry-count	10
hba0-enable-adapter- hard-loop	0
hba0-adapter-hard-loop-ID	0
hba0-enable-64bit-addressing	0
hba0-enable-LIP-reset	0
hba0-enable-LIP-full-login	1
hba0-enable-target-reset	0: non-clustered, =1: clustered
hba0-reset-delay	5
hba0-port-down-retry-count	30
hba0-link-down-error	1
hba0-loop-down-timeout	60

Parameters	Recommended settings
hba0-connection-options	1: switched fabric 2: point-to-point
hba0-device-configuration-mode	1
hba0-fc-tape	0
hba0-command-completion-option	1

In Table 5-8 we show the QLA2310F, QLA2340, and QLA2342 recommended settings.

Table 5-8 QLogic QLA2310F, QLA2340, QLA2342 recommendations

Parameters	Recommended settings
hba0-max-frame-length	2048
hba0-max-iocb-allocation	256
hba0-execution-throttle	31
hba0-login-timeout	4
hba0-login-retry-count	1
hba0-fabric-retry-count	10
hba0-adapter-hard-loop-ID	0
hba0-enable-64bit-addressing	0
hba0-enable-LIP-reset	0
hba0-enable-LIP-full-login	1
hba0-enable-target-reset	0: disabled
hba0-reset-delay	5
hba0-port-down-retry-count	30
hba0-link-down-error	1
hba0-loop-down-timeout	60
hba0-connection-options	0: loop only 1: point-to-point(fabric)
hba0-device-configuration-mode	1: use port name

Parameters	Recommended settings
hba0-fc-tape	0
hba0-fc-data-rate	2: auto-negotiate
hba0-command-completion-option	1
persistent binding only option	0: reports the discovery of persistent bound and non-bound devices to the operating system



Cisco directors and switches

The Cisco portfolio is based on multilayer network and storage intelligence. The switches allow for the configuration of scalable solutions that can help address the need for high performance and reliability in environments ranging from small workgroups to large, integrated enterprise SANs.

The MDS 9000 family introduces SAN capabilities, including Cisco's Virtual SAN (VSAN) technology, designed to enable efficient SAN utilization by creating hardware-based isolated environments within a single SAN fabric. Each VSAN can be zoned as a typical SAN and maintains its own fabric services for added scalability and resilience.

Overall, the MDS 9000 family is designed for investment protection, flexibility, scalability, robustness, advanced diagnostics, and integration between low cost workgroup, and enterprise SANs.

6.1 Overview

The Cisco Systems MDS 9216 (2062-D01) Multilayer Fabric Switch and MDS 9509 (2062-D07 / T07) Multilayer Director are now available from IBM and Authorized IBM Business Partners. The Cisco MDS 9000 family provides 1 to 2 Gb/s Fibre Channel capability and performance with advanced intelligence to help address the demands for security, performance, and manageability required to consolidate geographically dispersed SAN islands into a large enterprise SAN. They are designed to be fully interoperable with other members of the Cisco MDS 9000 family, and are based on a multilayer network and storage intelligence.

6.1.1 Cisco directors and switches

The following Cisco products are now available as part of the IBM Storage Networking solutions.

MDS 9509 Multilayer Director

The Cisco Systems MDS 9509 Model D07 (IBM 2062-D07) is a fourteen-rack-unit (14 RU) Fibre Channel director that can support from 32 to 224 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gb/s or 2 Gb/s Fibre Channel and are auto-sensing.

The chassis has nine slots, of which two are reserved for dual redundant supervisor modules. The dual supervisor modules provide the logic control for the director and also provide high availability and traffic load balancing capabilities across the director. Either supervisor module can control the whole director, with the standby supervisor module providing full redundancy in the event of an active supervisor failure.

The remaining seven slots can contain a mixture of switching modules which provide either 16 or 32 ports per module. Each module supports longwave or shortwave SFPs supporting the LC interface.

The director is configured with dual redundant power supplies either of which can supply power for the whole chassis. It also includes a hot-swappable fan tray with nine fans to manage the cooling and airflow for the entire director.

Shown in Figure 6-1 is the MDS 9509 Multilayer Director.



Figure 6-1 MDS 9509 Multilayer Director (IBM 2062-D07)

MDS 9509 Multilayer Director

The MDS 9509 Multilayer Director (IBM 2062-T07) product is designed for the Telecommunication industry and provides DC power supplies. We do not cover this specific model in this redbook.

MDS 9216 Multilayer Switch

The Cisco Systems MDS 9216 Model D01 (IBM 2062-D01) is a three-rack-unit (3 RU) fabric switch that can support from 16 to 48 shortwave or longwave SFP fiber optic transceivers. These ports fully support either 1 Gb/s or 2 Gb/s Fibre Channel and are auto-sensing.

The switch is configured with dual redundant power supplies either of which can supply power for the whole switch. It also includes a hot-swappable fan tray with four fans to manage the cooling and airflow for the entire switch.

The chassis consists of two slots, which contain these modules:

- ▶ **Supervisor module:** This provides the supervisor function for the switch and has 16 standard Fibre Channel ports
- ▶ **Switching module:** This provides an optional number of Fibre Channel ports and can be configured with either a 16 or 32 port switching module.

The 9216 switch shares a common architecture with the Cisco MDS 9500 series of multilayer directors, making it an intelligent and flexible fabric switch.

The MDS 9216 switch is shown in Figure 6-2.



Figure 6-2 MDS 9216 Multilayer Switch (IBM 2062-D01)

6.1.2 Supported attachments

Both the 9509 and 9216 support the following attachment types:

- ▶ FCP
- ▶ FC_AL (including public and private loop support)
- ▶ Cascading (attaching multiple switches or directors together)
- ▶ Interoperability

The unique architecture of the 9509 Multilayer Director allows integration of new transport protocols for greater flexibility. For example, the 9509 is designed to support Fibre Channel, Internet SCSI (iSCSI) and Fibre Channel over IP (FCIP).

IBM plans to offer an IP Storage Services Module feature, which would include Gigabit Ethernet, to enable Internet SCSI (iSCSI) and Fibre Channel over IP (FCIP) capabilities for the 9216.

6.1.3 Supported port-types

The ports provide an auto-sensing 1 or 2 Gb/s SFP that supports LC interfaces. The port modes include:

E_Port

In expansion port (E_Port) mode, an interface functions as a fabric expansion port. This port may be connected to another E_Port to create an ISL between two switches. E_Ports carry frames between switches for configuration and fabric management. They serve as a conduit between switches for frames destined to remote N_Ports and NL_Ports. E_Ports support class 2, class 3, and class F service.

An E_Port connected to another switch may also be configured to form a PortChannel.

F_Port

In fabric port (F_Port) mode, an interface functions as a fabric port. This port may be connected to a peripheral device (server or disk) operating as an N_Port. An F_Port can be attached to only one N_Port. F_Ports support class 2 and class 3 service.

FL_Port

In fabric loop port (FL_Port) mode, an interface functions as a fabric loop port. This port may be connected to one or more NL_Ports (including FL_Ports in other switches) to form a public arbitrated loop. If more than one FL_Port is detected on the arbitrated loop during initialization, only one FL_Port becomes operational and the other FL_Ports enter non-participating mode. FL_Ports support class 2 and class 3 service.

TL_Port

In translatable loop port (TL_Port) mode, an interface functions as a translatable loop port. It may be connected to one or more private loop devices (NL_Ports). TL_Port mode is specific to Cisco MDS 9000 family switches and have similar properties as FL_Ports. TL_Ports enable communication between private loop devices and one of the following target devices:

- ▶ A device attached to any switch on the fabric
- ▶ A device on a public loop anywhere in the fabric
- ▶ A device on a different private loop anywhere in the fabric
- ▶ A device on the same private loop

TL_Ports support class 2 and class 3 services.

TE_Port

In trunking E_Port (TE_Port) mode, an interface functions as a trunking expansion port. It may be connected to another TE_Port to create an Extended ISL (EISL) between two switches. TE_Ports are specific to Cisco MDS 9000 family switches. They expand the functionality of E_Ports to support these features:

- ▶ VSAN trunking
- ▶ Transport quality of service (QoS) parameters
- ▶ Fibre Channel trace (**fctrace**) feature

In TE_Port mode, all frames are transmitted in EISL frame format, which contains VSAN information. Interconnected switches use the VSAN ID to multiplex traffic from one or more VSANs across the same physical link. This feature is referred to as trunking in the Cisco MDS 9000 family.

TE_Ports support class 2, class 3, and class F service.

SD_Port

In SPAN destination port (SD_Port) mode, an interface functions as a switch port analyzer (SPAN). The SPAN feature is specific to switches in the Cisco MDS 9000 family. It monitors network traffic that passes through a Fibre Channel interface. This monitoring is done using a standard Fibre Channel analyzer (or a similar switch probe) that is attached to an SD_Port. SD_Ports do not receive frames, they merely transmit a copy of the source traffic. The SPAN feature is non-intrusive and does not affect switching of network traffic for any SPAN source ports.

Fx_Port

Interfaces configured as Fx_Ports are allowed to operate in either F_Port or FL_Port mode. The Fx_Port mode is determined during interface initialization depending on the attached N_Port or NL_Port. This administrative configuration disallows interfaces to operate in any other mode — for example, preventing an interface to connect to another switch.

Auto Mode

Interfaces configured in auto mode are allowed to operate in one of the following modes: F_Port, FL_Port, E_Port, or TE port. The port mode is determined during interface initialization. For example, if the interface is connected to a node (server or disk), it operates in F_Port or FL_Port mode depending on the N_Port or NL_Port mode. If the interface is attached to a third-party switch, it operates in E_Port mode. If the interface is attached to another switch in the Cisco MDS 9000 family, it may become operational in TE_Port mode. TL_Ports and SD ports are not determined during initialization and are administratively configured.

Shown in Figure 6-3 is an example of the port types available with the Cisco MDS 9000 family of products.

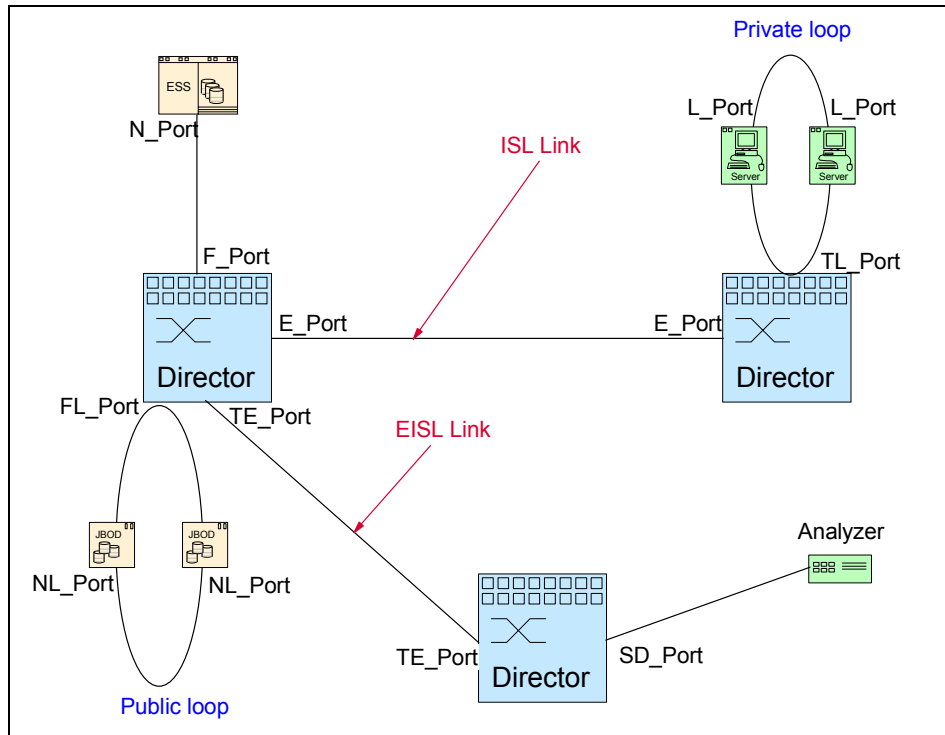


Figure 6-3 Cisco MDS 9000 family port types

6.1.4 Availability

The 9509 highlights director class management, serviceability, and availability, including hot swap support and redundancy of active hardware components, and non-disruptive software upgrade support on all active components. Each supervisor module is capable of providing the ability to automatically restart failed processes. In the event that a supervisor module is reset because of hardware failure or service action, complete synchronization between the active and standby supervisor modules support graceful process failover with no traffic disruption.

The director also provides management and software based availability features such as these:

- ▶ Non-disruptive software upgrade capability
- ▶ Failover of supervisor module code to the redundant supervisor module
- ▶ Protection against link failure through the use of PortChannels; this is discussed in “PortChannels” on page 210

- ▶ Non-disruptive restart of a failed process running on the same supervisor module

The 9216 product family includes the following features to help provide the highest reliability.

- ▶ Provides power-on self testing
- ▶ Detects errors, isolates faults and performs parity checking
- ▶ Remote diagnostics using Call Home features
- ▶ LED displays that summarize the status for the switching modules, supervisor modules, power supply assembly and fan assembly

6.1.5 Scalable capacity

The 9509 offers a scalable solution from 32 to 224 Fibre Channel ports. The director supports from 2 to 7 switching modules that provide either 16 or 32 ports per module for a total of 224 Fibre Channel ports in a single chassis.

The 9216 includes 16 Fibre Channel ports in the base module and can accept one additional 16 or 32 port module allowing for a total capacity of 16, 32 or 48 ports.

6.2 MDS 9509 Multilayer Director component overview

In the section below we will discuss the main components of the Cisco MDS 9509 Multilayer Director. Many of these components are common to both the 9509 and the 9216.

6.2.1 Chassis

The 9509 has a 9-slot chassis that is 14-RU high and can be installed inside a standard EIA rack providing from 32 to 224 ports. Of these 9-slots, 2 are reserved for the redundant supervisor modules with the remaining 7-slots available for switching modules (switching modules contain the actual FC ports or SFPs). The chassis also contains the dual power supplies which are located in the base of the chassis. The fan tray with redundant fans is located at the side of the chassis.

The layout of the 9509 is illustrated in Figure 6-4.

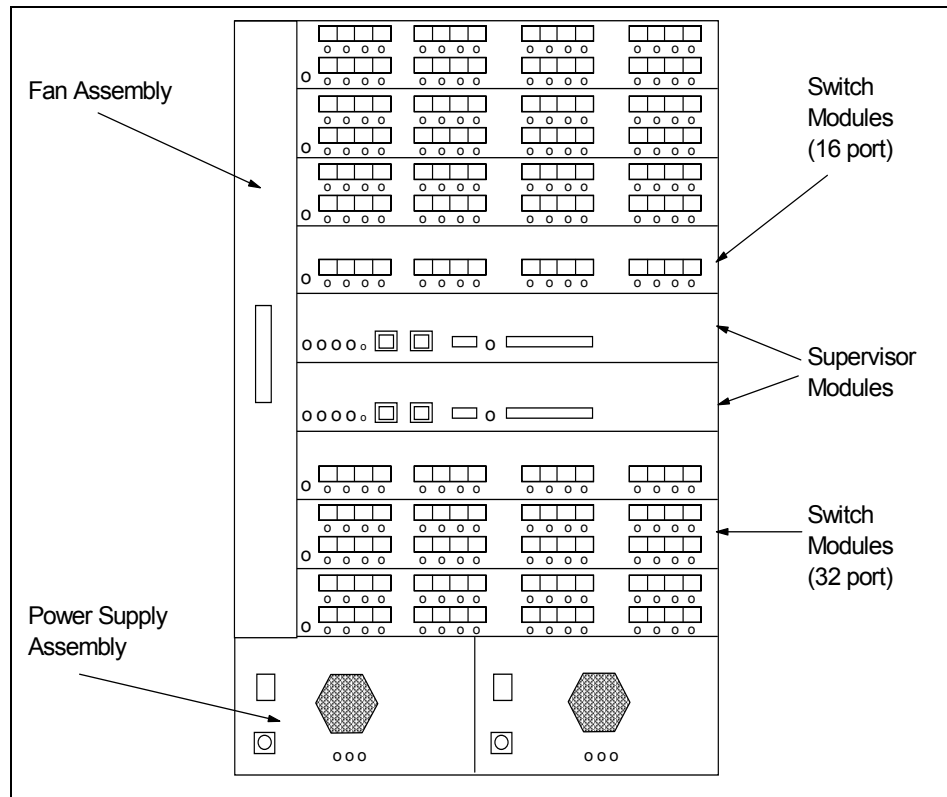


Figure 6-4 Cisco MDS 9509 Multilayer Director layout

6.2.2 Backplane

The backplane of the 9509 provides the connectivity for two supervisor modules and up to seven switching modules. As well as the supervisor and switching modules the redundant power supplies and the redundant dual clock modules also plug directly into the backplane.

If one clock module fails, the remaining clock module will take over operation of the director; however this will cause a reset of the director. Although there are dual redundant clock modules, if one needs to be replaced, then an outage will be required, as these modules are not currently hot-pluggable.

6.2.3 Power supply

The Cisco MDS 9509 provides dual power supplies as standard and either of these assemblies can be used to provide power to the entire chassis. The power supplies are hot-swappable and provide self-monitoring functions by reporting their status to the supervisor module.

The 9500 chassis has three power supply options. The DC power is for a specific environment, namely the telecommunication industry. When choosing a power supply, you need to consider the number of switching modules that will be installed in the director. For example, the 2500 W power supply, when connected to a 110 VAC power source, can only provide enough power to drive the redundant supervisor cards (220 W each), plus 3 to 4 switching modules (200-220 W each).

Listed in Table 6-1 are the power supply specifications.

Table 6-1 Power supply voltages

Power Supply	Voltage	Total Power	AC Plug Type
2500W AC	110 VAC 220 VAC	1300 2500	20 A
2500W DC	N/A	2500	
4000W AC	220 VAC	4000	30 A

To assist with configuring the correct power supply, in Table 6-2 we describe the power requirements for each of the supervisor and switching modules that can be installed into the MDS 9500 chassis. You can use this table to determine what power supply is required.

Table 6-2 Power consumption by module

Module Type	Power rating
Supervisor module	220 W
16 port switching module	220 W
32 port switching module	200 W

As specifications may change, it is wise to check at order time as to the power supply options that are available.

6.2.4 Fan assembly

The 9509 contains a hot-swappable fan assembly that has nine fans. The fan assembly provides 800 cubic feet per minute (cfm) of cooling, allowing for 410 W of power dissipation per slot.

Sensors on the supervisor module monitor the internal air temperature. If the air temperature exceeds a preset threshold, the environmental monitor displays warning messages.

If an individual fan (or fans) within the assembly fails, the Fan Status LED turns red. Individual fans cannot be replaced; however, the fan assembly can be replaced. It will continue to run if the fan assembly is removed, as long as preset temperature thresholds have not been exceeded. This allows you to swap out a fan assembly without having to bring the system down.

The director is designed with a side-to-side airflow, which is commonly used in LAN switching environments. Sufficient space between racks maybe required to provide adequate cooling.

6.2.5 Supervisor module

The supervisor module is the heart of the 9509 as it provides the control and management functions for the director, as well as an integrated crossbar switching fabric. The crossbar fabric provides up to 720 Gb/s full duplex switching capacity.

The 9509 comes standard with two supervisor modules for redundancy. In the event of a supervisor module failing, the surviving module takes over operation of the director.

Control and management

The supervisor module provides the following control and management features:

- ▶ Multiple paths to avoid a single point of failure.
- ▶ A redundant central arbiter that provides traffic control and access fairness.
- ▶ Non-disruptive restart of a single failing process on the same supervisor. A kernel service running on the supervisor module keeps track of the high availability policy of each process and issues a restart when a process fails. The type of restart issued is based on the process's capability:
 - Warm or stateful (state is preserved)
 - Cold or stateless (state is not preserved)

If the kernel service is unable to perform a warm restart of the process, it issues a cold restart.

- Nondisruptive switchover from the active supervisor to a redundant standby without loss of traffic.

If the supervisor module has to be restarted, then the secondary supervisor (which is continuously monitoring the primary) takes over. Switchover is non-revertive. Once a switchover has occurred and the failed supervisor has been replaced or restarted, operation does not switch back to the original primary supervisor, unless it is forced to switch back or unless another failure occurs.

Crossbar switching fabric

The 9509 supervisor module provides a crossbar switching fabric that connects all the switching modules. Single fabric configurations provide 720 Gb/s full duplex speed with an 80 Gb/s bandwidth per switching module.

Dual fabric configurations provide 1.4 Tb/s speed with a 160 Gb/s bandwidth per switching module. The 9509 supports redundancy if dual supervisors are installed.

Upon power up with slot 5 active the supervisors negotiate to determine which one is active and which is the standby supervisor. Each supervisor exchanges its own status, and updates the signal quality error (SQE) periodically. If the active supervisor becomes disabled, the standby supervisor switches over to the active supervisor. Dual supervisor modules provide dual crossbar switching fabrics for redundancy.

Processor

The processor on the supervisor module is a Pentium® III class processor and is configured with the options shown in Table 6-3.

Table 6-3 Processor memory options

Memory	Byte
DRAM	up to 4 GB
1 internal CompactFlash card ¹	512 MB
1 external CompactFlash slot ²	512 MB

Note 1: The card stores software images.

Note 2: The slot is for optional cards to store additional images, and for configurations, debugging and syslog information.

Interfaces

The supervisor module supports the following interfaces:

- ▶ **Console:** Has an RJ-45 connection that allows you to:
 - Configure the Cisco MDS 9500 from the CLI
 - Monitor network statistics and errors
 - Configure SNMP agent parameters
 - Download software updates or distribute software images residing in Flash memory to attached devices
- ▶ **10/100 Mgmt:** Has a 10/100-Mbps Ethernet interface with an RJ-45 connection that provides network management capabilities.
- ▶ **COM2:** Connects to an external serial communication device, such as an uninterruptible power supply (UPS).
- ▶ **CompactFlash slot:** Provides a slot for an optional CompactFlash card. The cards can be used for storing additional software images, and configuration, debugging, and syslog information.

Switching modules

The Cisco MDS 9000 family of switching modules consists of two options, the 16 port 1 to 2 Gb/s Fibre Channel switching module, and the 32 port 1 to 2 Gb/s Fibre Channel switching module.

The MDS 9216 accepts one optional switching module, supporting 16 to 48 Fibre Channel ports per chassis, while the MDS 9509 requires a minimum of two and supports up to a maximum of seven optional switch modules, supporting 32 to 224 Fibre Channel ports per chassis.

With switching modules, network architects can allocate bandwidth to meet specific application requirements while decreasing the switching footprint which can help lower the overall storage network deployment cost. Optional switching modules are designed to be interchanged or shared between all Cisco MDS 9000 Directors and Fabric Switches.

Attention: IBM has recently announced the following enhancements to the Cisco portfolio:

An 8-Port IP Line Card for iSCSI and FCIP (#2208): The optional 8-Port IP Services Line Card provides connection to Gigabit Ethernet networks and support for low-cost iSCSI host initiators to access Fibre Channel storage. The optional FCIP Activation features (#2209/#2210) allow the 8-Port IP Services Line Card to extend native Fibre Channel Fabric over LAN and WAN infrastructure.

The 8-Port IP Services Line Card has been designed to enable the use of iSCSI to extend the benefits of Fibre Channel SAN-based storage to IP-enabled servers at a lower cost point and to a broader array of workgroup platforms than supported with traditional SAN deployments. The iSCSI capability helps increase storage utilization and availability through consolidation of IP and Fibre Channel block storage, supporting transparent storage administration practices such as zoning storage resources.

The 8-Port IP Services Line Card may be ordered with either Short Wave (SW) SFP or Long Wave (LW) SFP optic transceivers; however, only one SFP is required as a minimum configuration. The 8-Port IP Services line card is compatible with Tri-Rate SFPs (#5210 and #5220) only.

FCIP Activation for 8-Port IP Services Line Card (#2209 and #2210): FCIP Activation on the 8-Port IP Services Line Card is designed to help simplify data protection and business continuance strategies by enabling backup, remote replication, and disaster recovery over LAN and WAN distances using open-standard FCIP tunneling. Specifically, FCIP enables support for extended distance Fibre Channel fabric operation, while supporting the Cisco MDS 9000 family of enhanced capabilities including VSANs, advanced traffic management, and security across LAN and WAN connections.

Additionally, FCIP improves utilization of WAN resources for backup and replication, by tunneling up to 3 virtual Inter Switch Links (ISLs) on a single Gigabit Ethernet port. Finally, FCIP can help reduce SAN complexity by eliminating the need to deploy and manage a separate remote connectivity platform.

FCIP Activation is available for the Cisco MDS 9216 Multilayer Fabric Switch with feature number 2209, and for the Cisco MDS 9509 Multilayer Director with feature 2210.

16 port switching module

The 16 port switching module is designed to deliver high performance for demanding storage networking applications. Autosensing 1 Gb/s and 2 Gb/s ports deliver up to 64 Gb/s of continuous aggregate bandwidth or 200 MB/s and 255 buffer credits per port.

Note: The 64 Gb/s continuous aggregate bandwidth is based on 2 Gb/s per port in full duplex mode.

That is: 16 ports @ 2 Gb/s (or 200 MB/s) in both directions = 64 Gb/s

The 16 port module is designed for attaching high-performance servers and storage subsystems as well as for connecting to other switches via ISL connections.

Figure 6-5 shows a picture of the 16 port switching module for the Cisco MDS 9000 family.



Figure 6-5 16 port switching module

32 port switching module

The 32 port switching module is designed to deliver an optimal balance of performance and port density. This module provides high line-card port density along with 32 Gb/s of total bandwidth and 12 buffer to buffer credits per port. Bandwidth is allocated across eight 4-port groups, providing 4 Gb/s (200 MB/s) of sustained bandwidth per port-group. This module provides a low-cost means of attaching lower performance servers and storage subsystems to high-performance crossbar switches without requiring ISLs.

Note: The 32 Gb/s continuous aggregate bandwidth is based on providing 4 Gb/s bandwidth to each port-group (each port-group contains 4-ports).

That is: 8 port-groups @ 4 Gb/s in both directions = 32 Gb/s

4 Gb/s to each port group does not mean that every port has only 1 Gb/s bandwidth. The whole port-group has an aggregate bandwidth of 4 Gb/s. One port could be transmitting at 2 Gb/s with the other ports sharing the remaining bandwidth.

By combining 16 and 32 port switching modules in a single, modular chassis administrators can configure price/performance-optimized storage networks for a wide range of application environments.

Switching modules are designed to be interchanged or shared between all Cisco MDS 9000 Directors and Fabric Switches

Figure 6-6 shows a picture of the 32 port switching module for the Cisco MDS 9000 family.

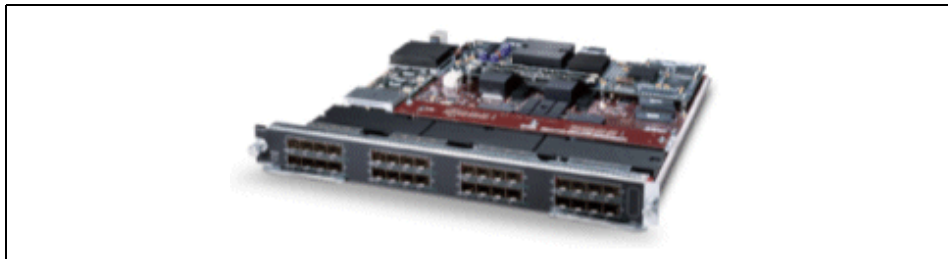


Figure 6-6 32 port switching module

Ports

The Cisco MDS 9000 family ports provides an auto-sensing 1 or 2 Gb/s SFP optic that support LC interfaces. The port modes include:

- ▶ **F_Port:** A fabric port (F_Port) connects the switch to a peripheral host or storage device. An F_Port can only be attached to one N_Port.
- ▶ **FL_Port:** A fabric loop port (FL_Port) connects the switch to a public loop.
- ▶ **TL_Port:** A translatable loop port (TL_Port) connects a switch to a private loop.
- ▶ **E_Port:** For connecting the switch to another switch or director. This uses an Inter Switch Link (ISL) to carry traffic from one switch to another.
- ▶ **TE_Port:** A Trunking E_Port for connecting a switch to multiple VSANs.

- **SD_Port:** A SPAN destination port acts as a snoopers port in a switched fabric environment. SPAN commands monitor the traffic entering and leaving the switch and copy the traffic to a specified destination (the SD_Port).

The ports are self-discovering when connected to the SAN and will automatically register as one of the types listed above.

SFP transceivers

The ports in the MDS 9000 series can be configured using a mixture of either shortwave (f/c 5230) or longwave (f/c 5240) SFP optic transceivers. Specific configuration options should be checked when ordering.

Attention: IBM has announced this feature enhancement to the Cisco portfolio:

The Tri-Rate SW SFP Transceiver (#5210) and Tri-Rate LW SFP Transceiver (#5220): The Cisco Tri-Rate SFP can run in three modes of operation: 1/2 Gb/s Fibre Channel as well as 1 Gb/s Ethernet, supporting a single SFP format across all MDS 9000 port types. The SW Tri-Rate SFP can be used for distances up to 500 meters. The LW Tri-Rate SFP can be used for distances up to 10 km.

Fibre Channel ports on the Cisco MDS 9216 Switch and MDS 9509 Director may be configured to accept either SW SFP or LW SFP optic transceivers; however, when ordering, all available Fibre Channel ports must be populated with SFP optic transceivers. Fibre Channel line cards are compatible with both Fibre Channel specific SFPs (#5230 and #5240) and Tri-Rate SFPs (#5210 and #5220).

Listed in Table 6-4 are the distance specifications for the SFPs.

Table 6-4 Distance specifications for SFP optics

Optics	Media	Supported distance
2 Gb/s - SW, LC SFP	50/125 micron multi-mode	300 m
2 Gb/s - SW, LC SFP	62.5/125 micron multi-mode	150 m
2 Gb/s - LW, LC SFP	9/125 micron single-mode	10 km

Buffer credits

Each port will support up to 255 buffer credits in the 16 port switching modules and up to 12 buffer credits per port for the 32 port switching module.

Physical dimensions

Listed in Table 6-5 are the physical specifications for a fully configured MDS 9509.

Table 6-5 Physical specifications for the MDS 9509

Cisco MDS 9509 Multilayer Director (IBM 2062-D07)	
Dimensions	62.3 cm H x 43.9 cm W x 46.8 cm D (24.5 in x 17.25 in x 18.4 in)
Rack Height	14U
Depth	55.0 cm (21.6 in)
Weight (fully configured chassis)	78 kg (170 lb)
Operating environment	
Temperature	0° to 40° C (32° to 104° C)
Relative Humidity	10% to 90%
Power Supplies	D07 Model: 2500 W AC
Input	D07 Model: 100 to 240 V AC 50-60 Hz nominal
Output	D07 Model: 1300 W at 100 to 120 V AC 2500 W at 200 to 240 V AC

6.3 MDS 9216 Multilayer Fabric Switch

In the following sections we discuss the main components of the Cisco MDS 9216 Multilayer Fabric Switch. Many of these components are common to both the 9509 and the 9216.

6.3.1 Chassis

The 9216 has a 2-slot chassis that is 3-RU high and can be installed inside a standard EIA rack providing 16, 32, or 48 ports. The chassis supports a supervisor module with an integrated 16 port switching module in slot 1 and an optional 16 port or 32 port switching module in slot 2.

Above the supervisor module in slot 1 is the interface module, which provides the local and remote management interfaces for the supervisor module. Slot 2 is reserved for the optional 16 or 32 port switching module, which enables additional ports to be installed for a maximum configuration of 48 ports. The chassis also contains the dual power supplies which are located in the rear of the chassis. The fan tray with redundant fans is located in the side of the chassis.

The layout of the MDS 9216 is illustrated in Figure 6-7.

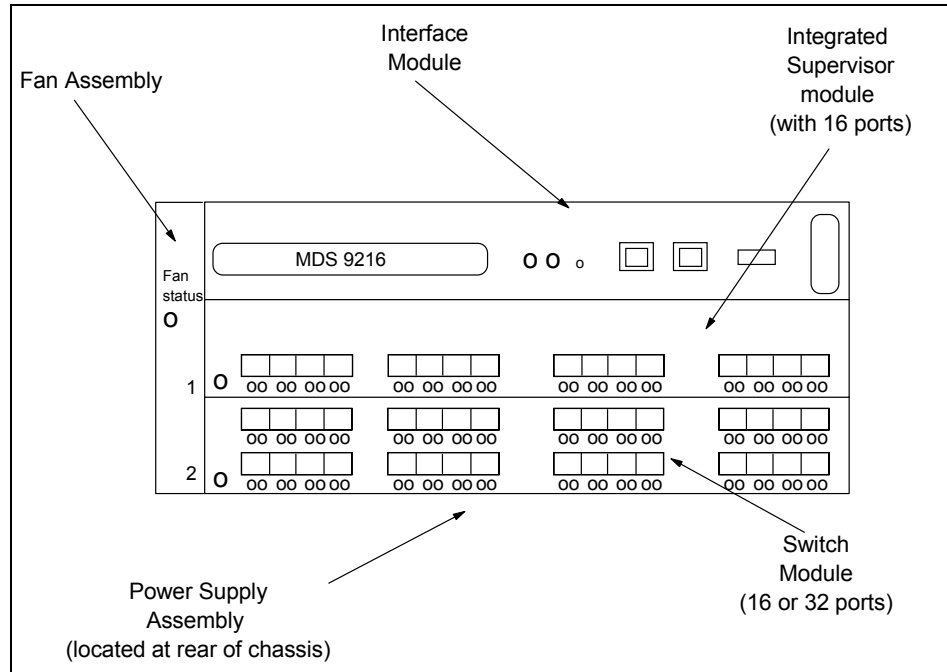


Figure 6-7 Cisco MDS 9216 Multilayer Fabric Switch layout

6.3.2 Backplane

The backplane of the 9216 provides the connectivity for one supervisor module and one switching module (either 16 or 32 port module). In addition to the supervisor and switching module, the redundant power supplies and the redundant dual clock modules also plug directly into the backplane.

If one clock module fails, the remaining clock module will take over operation of the director, but this will reset the switch. Although there are dual redundant clock modules, if one needs to be replaced, then an outage will be required, as these modules are not currently hot-pluggable.

6.3.3 Power supply

The 9216 supports dual 845 W AC power supplies that monitor output voltage and provide status to the supervisor module. Either power supply can provide power to the whole chassis in the event of a failure.

6.3.4 Fan assembly

The 9216 contains a hot-swappable fan assembly that has four fans. The fan assembly provides 270 cfm of cooling, allowing for 400 W of power dissipation per slot.

Sensors on the supervisor module monitor the internal air temperature. If the air temperature exceeds a preset threshold, the environmental monitor displays warning messages.

If an individual fan (or fans) within the assembly fails, the Fan Status LED turns red. Individual fans cannot be replaced; however, the fan assembly can be replaced. The 9216 continues to run if the fan assembly is removed, as long as preset temperature thresholds have not been exceeded. This allows you to swap out a fan assembly without having to bring the system down.

The switch is designed with side-to-side airflow, which is commonly used in LAN switching environments. Sufficient space between racks maybe required to provide adequate cooling.

6.3.5 Supervisor module

The supervisor module provides the control and management functions for the 9216, and it includes an integrated 16 port switching module. The supervisor module provides multiple communication and control paths to avoid a single point of failure, and a redundant central arbiter that provides traffic control and access fairness.

6.3.6 Processor

The supervisor module has a Pentium III class processor. It provides up to 2 GB of DRAM and has one internal CompactFlash card that provides 256 MB of storage for software images.

6.3.7 Interfaces

The interface module, which is located above slot 1, provides the following local and remote management interfaces for the supervisor module:

- ▶ **Console:** This has an RJ-45 connection that allows you to do these tasks:
 - Configure the switch from the CLI.
 - Monitor network statistics and errors.
 - Configure SNMP agent parameters.
 - Download software updates to the switch or distribute software images residing in Flash memory to attached devices.
- ▶ **10/100 MGMT:** This has a 10/100-Mbps Ethernet interface with an RJ-45 connection that provides network management capabilities.
- ▶ **COM2:** This connects to an external serial communication device, such as an uninterruptible power supply (UPS).

6.3.8 Switching modules

The 9000 family of switching modules consists of two options, the 16 port, 1 to 2 Gb/s Fibre Channel switching module, and the 32 port, 1 to 2 Gb/s Fibre Channel switching module.

The MDS 9216 accepts one optional switching module, supporting 16 to 48 Fibre Channel ports per chassis.

With switching modules, network architects can allocate bandwidth to meet specific application requirements while decreasing the switching footprint which can help lower the overall storage network deployment cost. Optional switching modules are designed to be interchanged or shared between all Cisco MDS 9000 Directors and Fabric Switches.

Attention: IBM has recently announced the following enhancements to the Cisco portfolio:

An 8-Port IP Line Card for iSCSI and FCIP (#2208): The optional 8-Port IP Services Line Card provides connection to Gigabit Ethernet networks and support for low-cost iSCSI host initiators to access Fibre Channel storage. The optional FCIP Activation features (#2209/#2210) allow the 8-Port IP Services Line Card to extend native Fibre Channel Fabric over LAN and WAN infrastructure.

The 8-Port IP Services Line Card has been designed to enable the use of iSCSI to extend the benefits of Fibre Channel SAN-based storage to IP-enabled servers at a lower cost point and to a broader array of workgroup platforms than supported with traditional SAN deployments. The iSCSI capability helps increase storage utilization and availability through consolidation of IP and Fibre Channel block storage, supporting transparent storage administration practices such as zoning storage resources.

The 8-Port IP Services Line Card may be ordered with either Short Wave (SW) SFP or Long Wave (LW) SFP optic transceivers; however, only one SFP is required as a minimum configuration. The 8-Port IP Services line card is compatible with Tri-Rate SFPs (#5210 and #5220) only.

FCIP Activation for 8-Port IP Services Line Card (#2209 and #2210): FCIP Activation on the 8-Port IP Services Line Card is designed to help simplify data protection and business continuance strategies by enabling backup, remote replication, and disaster recovery over LAN and WAN distances using open-standard FCIP tunneling. Specifically, FCIP enables support for extended distance Fibre Channel fabric operation, while supporting the Cisco MDS 9000 family of enhanced capabilities including VSANs, advanced traffic management, and security across LAN and WAN connections.

Additionally, FCIP improves utilization of WAN resources for backup and replication, by tunneling up to 3 virtual Inter Switch Links (ISLs) on a single Gigabit Ethernet port. Finally, FCIP can help reduce SAN complexity by eliminating the need to deploy and manage a separate remote connectivity platform.

FCIP Activation is available for the Cisco MDS 9216 Multilayer Fabric Switch with feature number 2209, and for the Cisco MDS 9509 Multilayer Director with feature 2210.

16 port switching module

The 16 port switching module is designed to deliver high performance for demanding storage networking applications. Autosensing 1 Gb/s and 2Gb/s ports deliver up to 64 Gb/s of continuous aggregate bandwidth or 200 MB/s and 255 buffer credits per port.

Note: The 64 Gb/s continuous aggregate bandwidth is based on 2 Gb/s per port in full duplex mode.

That is: 16 ports @ 2 Gb/s (or 200 MB/s) in both directions = 64 Gb/s.

The 16 port module is designed for attaching high-performance servers and storage subsystems as well as for connecting to other switches via ISL connections.

Figure 6-8 shows the 16 port switching module for the Cisco MDS 9000 family.

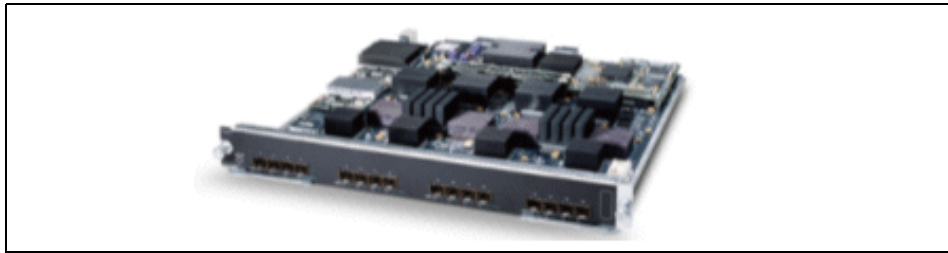


Figure 6-8 16 port switching module

32 port switching module

The 32 port switching module is designed to deliver an optimal balance of performance and port density. This module provides high line-card port density along with 32 Gb/s of total bandwidth and 12 buffer to buffer credits per port. Bandwidth is allocated across eight 4-port groups, providing 4 Gb/s (200 MB/s) of sustained bandwidth per port group. This module provides a low-cost means of attaching lower performance servers and storage subsystems to a high performance crossbar switch without requiring ISLs.

Note: The 32 Gb/s continuous aggregate bandwidth is based on providing 4 Gb/s bandwidth to each port-group (each port-group contains 4-ports).

That is: 8 port-groups @ 4 Gb/s in both directions = 32 Gb/s.

Notice that 4 Gb/s to each port group does not mean that every port has only 1 Gb/s bandwidth. The whole port-group has an aggregate bandwidth of 4 Gb/s. One port could be transmitting at 2 Gb/s, with the other ports sharing the remaining bandwidth.

By combining 16 and 32 port switching modules in a single, modular chassis, administrators can configure price/performance optimized storage networks for a wide range of application environments.

Switching modules are designed to be interchanged or shared between all Cisco MDS 9000 Directors and Fabric Switches

Figure 6-9 shows a picture of the 32 port switching module for the Cisco MDS 9000 family.

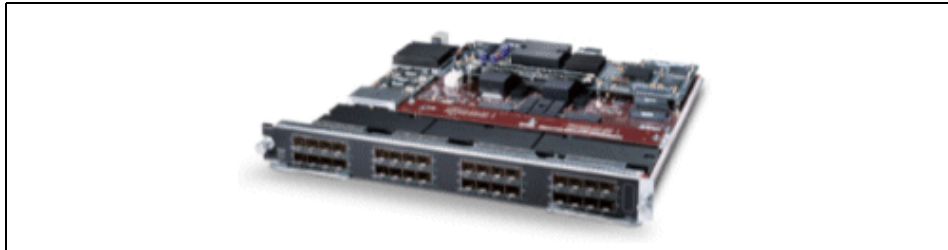


Figure 6-9 32 port switching module

Ports

The ports provide an auto-sensing 1 or 2 Gb/s SFP that support LC interfaces. The port modes include:

- ▶ **F_Port:** A fabric port (F_Port) connects the switch to a peripheral host or storage device. An F_Port can only be attached to one N_Port.
- ▶ **FL_Port:** A fabric loop port (FL_Port) connects the switch to a public loop.
- ▶ **TL_Port:** A translatable loop port (TL_Port) connects a switch to a private loop.
- ▶ **E_Port:** For connecting the switch to another switch or director. This uses an Inter Switch Link (ISL) to carry traffic from one switch to another.
- ▶ **TE_Port:** This is a Trunking E_Port for connecting a switch to multiple VSANs.

- **SD_Port:** A SPAN destination port acts as a snooper port in a switched fabric environment. SPAN commands monitor the traffic entering and leaving the switch and copy the traffic to a specified destination (the SD_Port).

The ports are self-discovering when connected to the SAN and will automatically register as one of the types listed above.

SFP transceivers

The ports in the MDS 9000 series can be configured using a mixture of either shortwave (f/c 5230) or longwave (f/c 5240) SFP optic transceivers. Configuration options should be checked when ordering.

Attention: IBM has announced the following feature enhancement to the Cisco portfolio:

The Tri-Rate SW SFP Transceiver (#5210) and Tri-Rate LW SFP Transceiver (#5220): The Cisco Tri-Rate SFP can run in three modes of operation: 1/2 Gb/s Fibre Channel as well as 1 Gb/s Ethernet, supporting a single SFP format across all MDS 9000 port types. The SW Tri-Rate SFP can be used for distances up to 500 meters. The LW Tri-Rate SFP can be used for distances up to 10 km.

Fibre Channel ports on the Cisco MDS 9216 Switch and MDS 9509 Director may be configured to accept either SW SFP or LW SFP optic transceivers; however, when ordering, all available Fibre Channel ports must be populated with SFP optic transceivers. Fibre Channel line cards are compatible with both Fibre Channel specific SFPs (#5230 and #5240) and Tri-Rate SFPs (#5210 and #5220).

Listed in Table 6-6 are the distance specifications for the SFPs.

Table 6-6 Distance specifications for SFP optics

Optics	Media	Supported distance
2 Gb/s - SW, LC SFP	50/125 micron multi-mode	300 m
2 Gb/s - SW, LC SFP	62.5/125 micron multi-mode	150 m
2 Gb/s - LW, LC SFP	9/125 micron single-mode	10 km

Buffer credits

Each port will support up to 255 buffer credits in the 16 port switching modules and up to 12 buffer credits per port for the 32 port switching module.

Physical dimensions

Listed in Table 6-7 are the physical specifications for the MDS 9216.

Table 6-7 Physical specifications for the MDS 9216

Cisco MDS 9216 Multilayer Director (IBM 2062-D01)	
Dimensions	13.3 cm H x 43.9 cm W x 57.6 cm D (5.25 in x 17.25 in x 22.7 in)
Rack Height	3U
Weight (fully configured chassis)	32 kg (170 lb)
Operating environment	
Temperature	0° to 40° C (32° to 104° C)
Relative Humidity	10% to 90%
Power Supplies	845 W AC
Input	D07 Model: 100 to 240 V AC 50-60 Hz nominal

6.4 Cisco SAN capabilities

The following SAN capabilities are integrated into the base of all models within the MDS 9000 family, and are designed to simplify deployment and efficient operation of large-scale fabric installations.

6.4.1 PortChannels

PortChannels allow users to aggregate up to sixteen physical ISLs into a single logical bundle, designed to optimize bandwidth utilization and resilience between switches. The group of Fibre Channel ISLs designated to act as a PortChannel may consist of any port on any 16 Port switching module within the MDS 9000 chassis, allowing the overall PortChannel to remain active upon failure of one or more ports, or failure of one or more switching modules. These PortChannels:

- ▶ Increase the aggregate bandwidth on an ISL or EISL by distributing traffic among all functional links in the channel.
- ▶ Load balance across multiple links and maintains optimum bandwidth utilization. Load balancing is based on a source ID (SID), destination ID (DID), and optionally the originator exchange ID (OX ID) that identify the flow of the frame.

- Provide high availability on an ISL. If one link fails, traffic previously carried on this link is switched to the remaining links. If a link goes down in a PortChannel, the upper protocol is not aware of it. To the upper protocol, the link is still there, although the bandwidth is diminished. The routing tables are not affected by link failure. PortChannels can contain up to 16 physical links and can span multiple modules for added high availability.

Shown in Figure 6-10 is a diagram representing ISLs and PortChanneling.

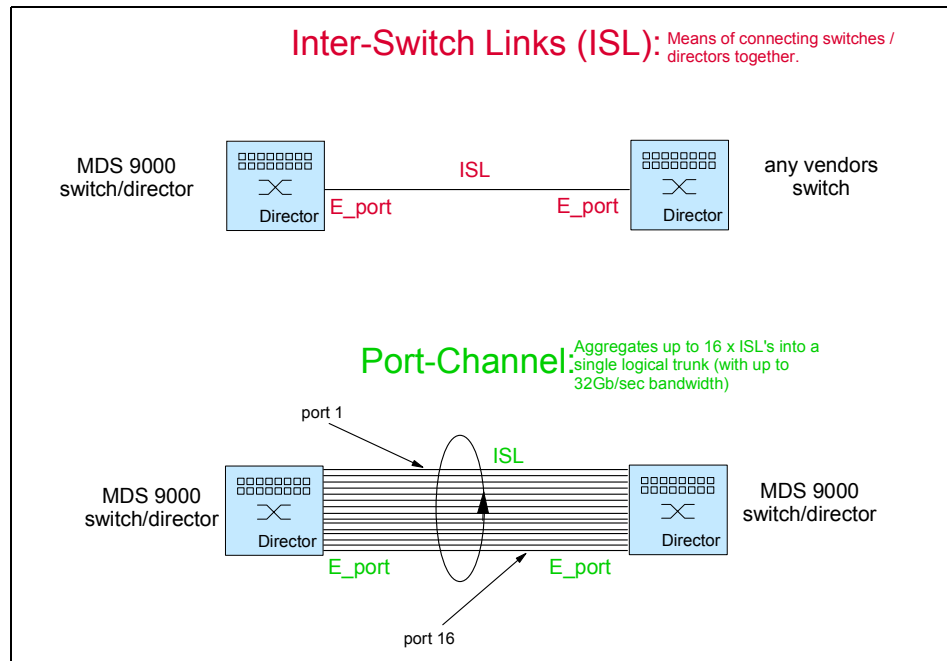


Figure 6-10 PortChannels and ISLs on the Cisco MDS 9000 family of switches

6.4.2 Trunking

The Cisco MDS 9000 family uses the term *trunking* to refer to an ISL link that carries one or more VSANs. Trunking ports receive and transmit EISL frames. EISL frames carry an EISL header containing the VSAN information. Once EISL is enabled on an E_Port, that port becomes a TE_Port.

Trunking is also referred to as VSAN Trunking, as it only applies to a VSAN. If a trunking enabled E_Port is connected to another vendor's switch, the trunking protocol ensures that the port will operate as a standard E_Port.

Shown in Figure 6-11 is a diagram of trunking. This also demonstrates how a combination of PortChannels and trunking can be used to create an aggregate bandwidth of up to 32 Gb/s between switches.

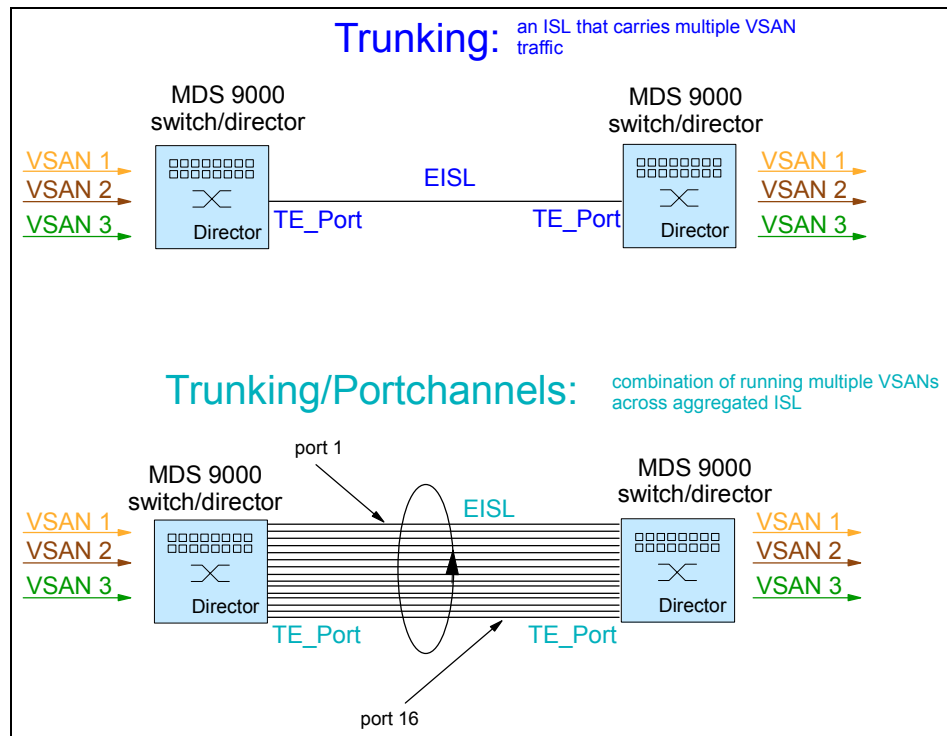


Figure 6-11 Trunking and PortChanneling

6.4.3 IP services

Switches in the Cisco MDS 9000 family support the following IP services:

- ▶ **IP over Ethernet:** These services are limited to management traffic.
- ▶ **IP over Fibre Channel (IPFC):** IPFC (RFC 2625) specifies how IP packets are transported using encapsulation schemes. By encapsulating IP frames into Fibre Channel frames, management information is exchanged among switches without requiring a separate Ethernet connection to each switch. Each switch includes:
 - Encapsulation for IP and Address Resolution Protocol (ARP) over Fibre Channel.
 - Address resolution uses the ARP server.

- ▶ **IP routing services:** These services include:
 - Ethernet or TCP/IP connection
 - Static IP routing services to enable management traffic between VSANs
 - DNS client support
 - Network Time Protocol (NTP) server that synchronizes the system clocks of network devices

6.4.4 Quality of Service

The switches provide low and high priority quality of service (QoS) queues. While time-critical traffic is marked as high priority traffic, all other traffic is assigned to the default low priority queue.

Quality of Service offers the following advantages:

- ▶ Guarantees absolute and relative bandwidth choices
- ▶ Provides latency to reduce frame loss
- ▶ Prioritizes transactional traffic over bulk traffic
- ▶ Supports multiple VSANs on the same fabric by guaranteeing bandwidth and latency available to each VSAN

The Cisco MDS 9000 family supports QoS for internally and externally generated control traffic. Within a switch the control traffic is sourced to the supervisor module and is treated as a high priority frame. A high priority status provides absolute priority over all other traffic and is assigned in the following cases:

- ▶ Internally generated time-critical control traffic (generally Class F frames).
- ▶ Externally generated time-critical control traffic entering a switch in the MDS 9000 range from another vendor's switch. High priority frames originating from other vendor switches retain the priority as they enter a switch in the MDS 9000 family.

By default, the QoS feature for control traffic is enabled but can be disabled if required.

Congestion control

Forward Congestion Control (FCC) is a Cisco proprietary flow control mechanism that alleviates congestion on Fibre Channel networks. FCC reduces the congestion in the fabric without interfering with the standard Fibre Channel protocols. The FCC protocols increase the granularity and the scale of congestion control applied to any class of traffic.

Any switch in the network can detect congestion for an output port. The switches sample frames from the congested queue and generate messages about the congestion level upstream toward the source of the congestion.

The switch closest to the source, with FCC enabled, performs one of two actions:

- ▶ Forwards the frames as other vendor switches do.
- ▶ Limits the flow of frames from the port causing the congestion.

This is illustrated in Figure 6-12.

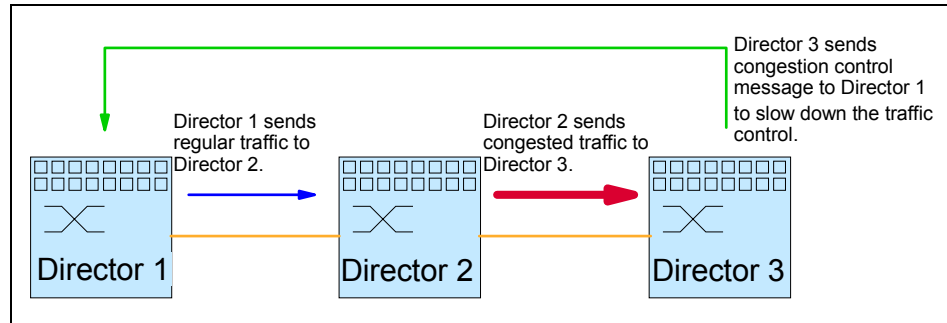


Figure 6-12 Forward Congestion Control

By default, the FCC protocol is disabled. You can enable the protocol globally for all the VSANs configured in the switch, or selectively enable or disable it for each VSAN.

Congestion control methods

With FCC enabled, these are the different congestion control methods:

- ▶ **Path quench control:** Reduces severe congestion temporarily by slowing the source to the whole path in the fabric.
- ▶ **Edge quench control:** Provides feedback to the source about the rate at which frames should be entered into the network (frame intervals).

FCC process

When a node in the network detects congestion for an output port, it generates an edge or a path quench message. These frames are identified by the Fibre Channel destination ID (DID) and the source ID (SID). A switch from other vendors simply forwards these frames.

Any receiving switch in the Cisco MDS 9000 family handles frames in one of these ways:

- ▶ It forwards the frame.
- ▶ It limits the rate of the frame flow in the congested port.

Behavior of the flow control mechanism differs, based on the Fibre Channel DID:

- ▶ If the Fibre Channel DID is directly connected to one of the switch ports, the input rate limit is applied to that port.
- ▶ If the destination of the edge quest frame is a Cisco domain or the next hop is a Cisco MDS 9000 family switch, the frame is forwarded.
- ▶ If neither of these conditions is true, then the frame is processed in the port going towards the FC DID.

All switches (including the edge switch) along the congested path process path quest frames. However, only the edge switch processes edge quest frames. The FCC protocol is implemented for each VSAN and can be enabled or disabled on a specified VSAN or for all VSANs at the same time.

Call home

The Cisco MDS 9000 family includes a call home facility that detects any failure alerts and forwards them with the relevant information via e-mail.

6.5 Software management

The Cisco MDS 9000 family provides three principal modes of management:

- ▶ **The MDS 9000 family command line interface (CLI):** This presents the user with a consistent, logical CLI, which adheres to the syntax of the widely known Cisco IOS CLI. This is an easy-to-use command interface which has broad functionality.
- ▶ **Cisco Fabric Manager:** This is a responsive, easy-to-use Java application that simplifies management across multiple switches and fabrics. It enables administrators to perform vital tasks such as topology discovery, fabric configuration and verification, provisioning, monitoring, and fault resolution. All functions are available through an interface that enables remote management from any location.
- ▶ **Integration with third-party storage management tools:** This may be used independently or in conjunction with third party management applications. Cisco provides an extensive application programming interface (API) for integration with third-party and user developed management tools

6.5.1 Cisco MDS 9000 Fabric Manager

The Cisco Fabric Manager is included with the Cisco MDS 9000 family of switches and is a Java and SNMP-based network fabric and device management tool. It comes with a GUI that displays real-time views of your network fabric and installed devices.

The Cisco Fabric Manager provides three views for managing your network fabric:

- ▶ The Device View displays a continuously updated physical picture of device configuration and performance conditions for a single switch.
- ▶ The Fabric View displays a view of your network fabric, including multiple switches.
- ▶ The Summary View presents a summary view of switches, hosts, storage subsystems, and VSANs.
- ▶ The Cisco Fabric Manager provides an alternative to the CLI for most switch configuration commands.

The Cisco Fabric Manager is included with each switch in the Cisco MDS 9000 family.

6.5.2 In-band management and out-of-band management

The Cisco Fabric Manager requires an out-of-band (Ethernet) connection to at least one Cisco MDS 9000 family switch to enable it to discover and manage the fabric.

The interface used for an out-of-band management connection is a 10/100 Mb/s Ethernet interface on the supervisor module, labeled mgmt0. The mgmt0 connection can be connected to a management network to access the switch through IP over Ethernet.

Ethernet connectivity is required to at least one Cisco MDS 9000 family switch. This connection is then used to manage the other switches using in-band (Fibre Channel) connectivity. Otherwise, you need to connect the mgmt0 port on each switch to your Ethernet network.

Each supervisor module has its own Ethernet connection; however, the two Ethernet connections in a redundant supervisor system operate in active or standby mode. The active supervisor module also hosts the active mgmt0 connection. When a failover event occurs to the standby supervisor module, the IP address and media access control (MAC) address of the active Ethernet connection are moved to the standby Ethernet connection. This eliminates any need for the management stations to relearn the location of the switch.

An example of an out-of-band management solution is shown in Figure 6-13.

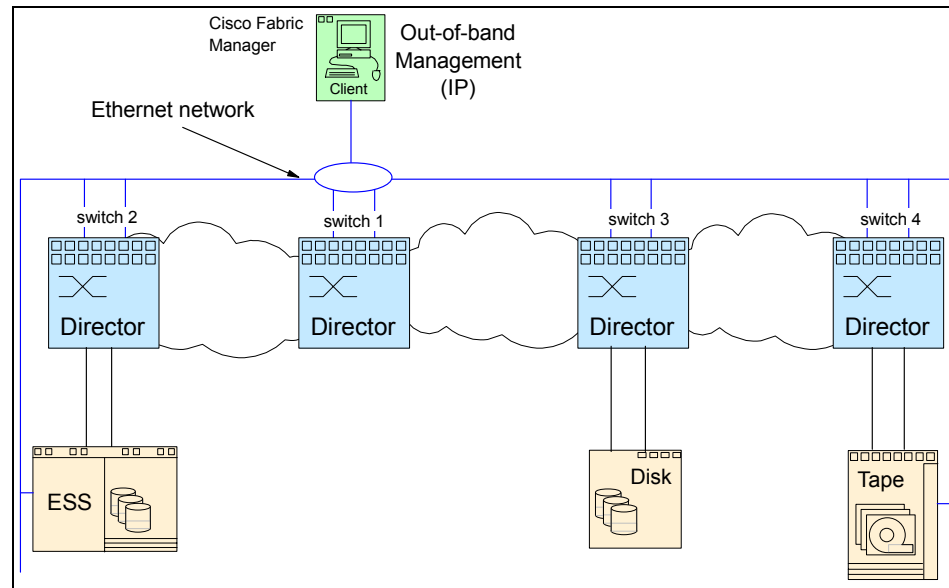


Figure 6-13 Out-of-band management connection

You can also manage switches on a Fibre Channel network using an in-band connection to the supervisor module. This in-band connection supports either management protocols over Fibre Channel or IP embedded within Fibre Channel. The Cisco MDS 9000 family supports RFC 2625 IP over Fibre Channel (IPFC), which allows IP to be transported between Fibre Channel devices over the Fibre Channel protocol, as shown in Figure 6-14.

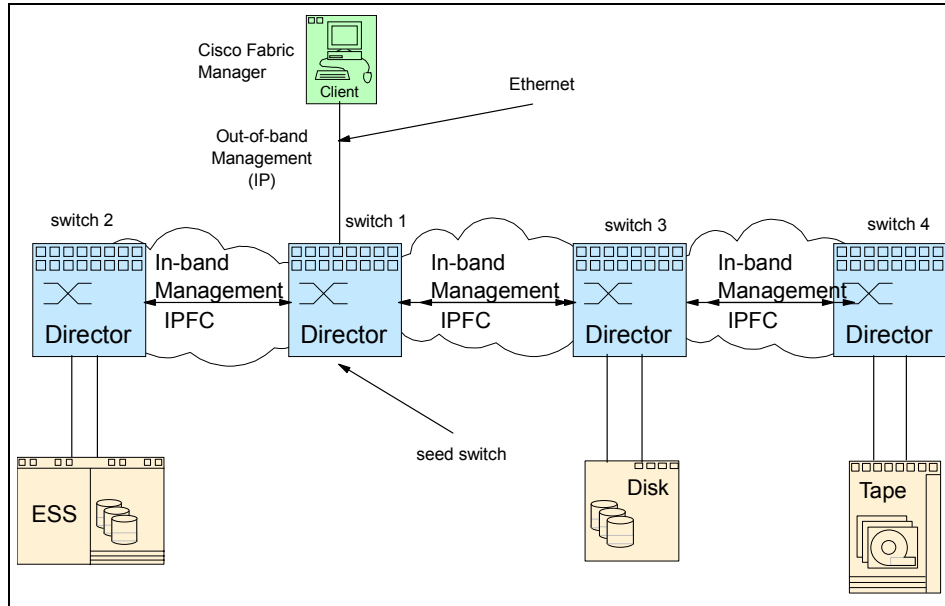


Figure 6-14 In-band management connection

IPFC encapsulates IP packets into Fibre Channel frames so that management information can cross the Fibre Channel network without requiring a dedicated Ethernet connection to each switch. IP addresses are resolved to the Fibre Channel address through Address Resolution Protocol (ARP). With host bus adapters (HBAs) that support IP drivers, this capability allows for a completely in-band management network. The switch also uses the in-band interface to discover its own environment, including directly connected and fabric-wide elements.

6.5.3 Using the setup routine

When you connect to a Cisco MDS 9000 family switch using the local console and start the switch for the first time, the system displays a setup routine that helps you perform the basic configuration required to manage and connect the switch to end nodes or other switches. The setup routine must be completed before you can connect to the switch or manage it using the Cisco Fabric Manager.

The setup routine prompts for the following configuration values:

- ▶ Default switch port status (enabled or disabled)
- ▶ Trunking (enabled or disabled)
- ▶ Zone policy (deny or permit access)

- ▶ IP address and subnet mask for mgmt0 (user supplied)
- ▶ Telnet/SSH policy (enabled or disabled)
- ▶ SNMPv3 user name and password (admin/admin_Pass)

In addition to these settings, each Cisco MDS 9000 family switch is configured with the following default values:

- ▶ VSAN membership: All ports are in VSAN 1
- ▶ Switch port speed and type: Autosense

6.5.4 Controlling administrator access with users and roles

The Cisco MDS 9000 family switches support role-based management access whether using the CLI or the Cisco Fabric Manager. This lets you assign specific management privileges to particular roles and then assign one or more users to each role.

When using the CLI, you can configure a local database of users using the CLI, or establish this database using a RADIUS server and then configure the CLI to verify console access with the RADIUS server. You can also use the Cisco Fabric Manager to configure RADIUS server for authenticating CLI access.

Cisco Fabric Manager uses SNMPv3 to establish role-based management access. After completing the setup routine, a single role, user name, and password are established. If the defaults are accepted during the setup routine, the user name is admin and the default password is admin_Pass. The role assigned to this user allows the highest level of privileges, which includes creating new users and roles. Use the Cisco Fabric Manager to create roles, and users, and to assign passwords as required for secure management access in your network.

6.5.5 Accessing Cisco Fabric Manager

Before you can access the Cisco Fabric Manager, you must complete the following tasks:

- ▶ A supervisor module must be installed on each switch that you want to manage
- ▶ The supervisor module must be configured with the following values using the setup routine or the CLI:
 - IP address assigned to the mgmt0 interface
 - SNMPv3 user name and password

6.5.6 Connecting to a supervisor module

The Cisco Fabric Manager software executables reside on each supervisor module of each Cisco MDS 9000 family switch in your network. The supervisor module provides an HTTP server that responds to browser requests and distributes the software to Windows or UNIX network management stations.

To install the software for the first time, or if you want to update or reinstall the software, access the supervisor module with a Web browser. When you click the install buttons on the Web page that is displayed, the software running on your workstation is verified to make sure you are running the most current version of the software. If it is not current, the most recent version is downloaded and installed on your workstation.

Although this redbook is not intended to take the place of an installation guide, we show some of the screens that are associated with the install of the product. We also point out some features of the software.

6.5.7 Software download and install

To download and install the software on your workstation, follow these steps:

- Enter the IP address or host name of the supervisor module in the address or location field of your browser. The supervisor module HTTP server displays the window, as shown in Figure 6-15.

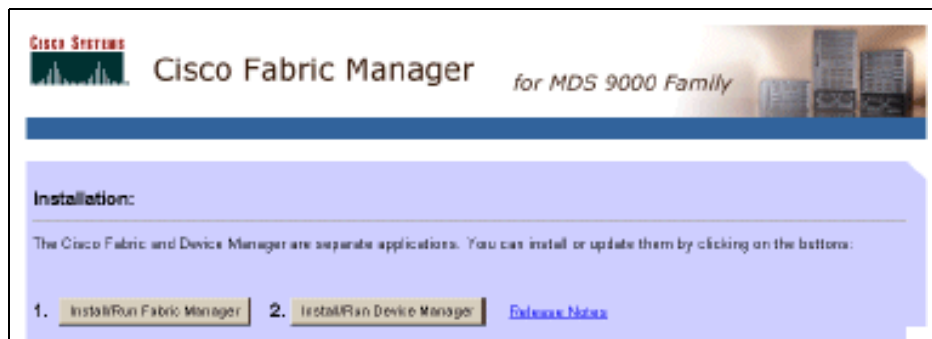


Figure 6-15 Cisco MDS 9000 family - Supervisor Web page

When you connect to the server for the first time, it checks to see if you have the correct Sun Java Virtual machine version installed on your workstation. If not, a link is provided to the appropriate Web page on Sun Microsystems's Web site so you can install it.

- Click the link to the Sun Java Virtual Machine software (if required) and install the software.

Using the instructions provided by the Sun Microsystems Web site to reconnect to the supervisor module by reentering the IP address or host name in the Location or Address field of your browser.

- Click either installation link (**Install Fabric Manager** or **Install Device Manager**).

You will see a prompt asking for permission to install the Java applets on your workstation, as shown in Figure 6-16.

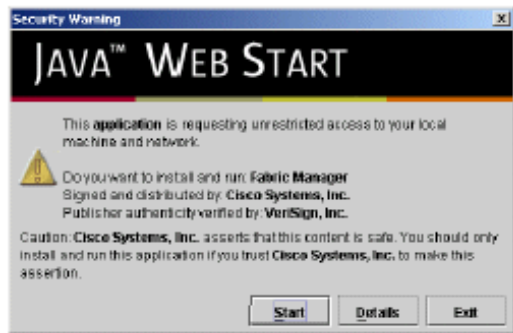


Figure 6-16 Java Web start warning

- Click **Start** to begin installing the software.

The Java Web Start application is automatically downloaded and installed on your workstation. Once the installation is complete, you can start the Cisco Fabric Manager directly from the Fabric Manager icon or the Device Manager icon on your desktop, or from the options on the Windows Start menu.

6.5.8 Launching views

To launch the Fabric Manager (Fabric View) or Fabric Device Manager (Device View and Summary View), follow these steps:

- Double-click the Fabric Manager icon or the Fabric Device Manager icon on your desktop or select the option from the Windows Start menu.

You will see the login screen, as shown in Figure 6-17.

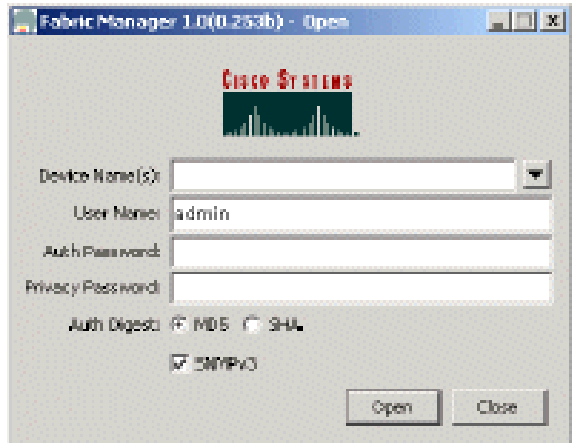


Figure 6-17 Fabric Manager - login screen

- ▶ Enter the IP address or device name in the Device Name(s) field.
Click the drop-down arrow to the right of the Device Name(s) field to display a list of devices that have been accessed previously.
- ▶ Enter a username and password.
- ▶ Enter the Privacy Password used for encrypting management traffic if the SNMPv3 Privacy option is enabled.
The Privacy option causes all management data to be encrypted. With SNMPv3, user names and passwords are always encrypted.
- ▶ Check the SNMPv3 check box to select SNMP version 3.
- ▶ Click **Open**.
You will see either the Fabric View or the Device View.

6.5.9 Discovering and viewing the network fabric

Cisco Fabric Manager automatically discovers all connected devices supporting the Fibre Channel GS3 (FC-GS3) standard, and all directly connected Fibre Channel devices. When you start the Fabric Manager, you enter the IP address (or DNS host name) of a seed switch. The Fabric Manager discovery process sends a Fibre Channel Service (FCS) request to each virtual SAN (VSAN) configured on the switch. VSANs are a unique feature of the Cisco MDS 9000 family switches. They allow multiple virtual SANs to be created within a single physical fabric.

After discovering all of the FC-GS3 devices in each VSAN connected to the seed switch, the discovery process sends SNMP requests over the Fibre Channel network to each discovered device to determine if other VSANs are configured elsewhere in the network fabric.

The discovery process is illustrated in Figure 6-18.

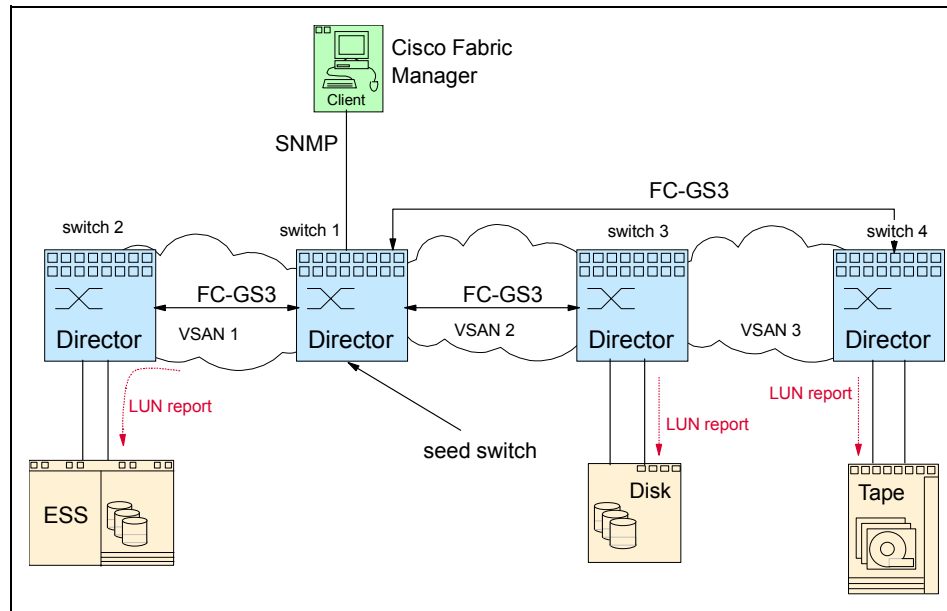


Figure 6-18 Cisco MDS 9000 Fabric Manager - discovery process

In addition to discovering devices that support FC-GS3, the seed switch discovers all directly connected FC devices. A separate discovery process is provided when a LUN report is requested for the first time. When this discovery process is initiated, the switch communicates with all the SCSI devices that are discovered to obtain information about logical unit numbers (LUNs) on connected storage devices. However, the discovery process will not discover any hosts or storage connected to the network fabric through a device unless it supports FC-GS3.

After you start Fabric Manager and discovery completes, you will see the Fabric View, as shown in Figure 6-19.

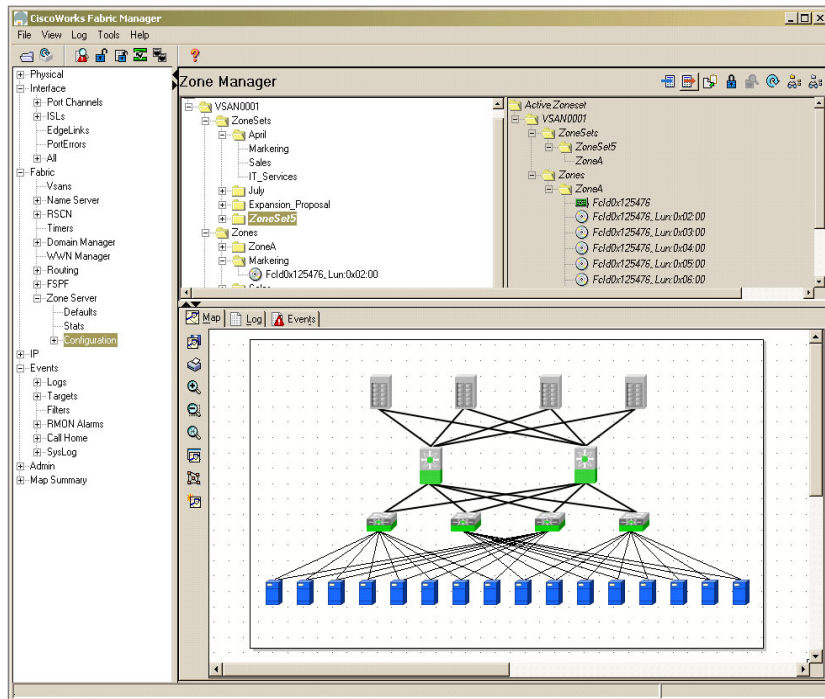


Figure 6-19 Fabric Manager - fabric view

The Fabric View provides a view of your network fabric, including all discovered switches, hosts, and storage devices.

You use the Fabric View to discover and view your fabric topology and to manage zones and zone sets. It is also convenient to use the Fabric View to manage other kinds of configuration involving more than one switch, such as VSANs and PortChannels. The following are some of the main fabric management tasks that you can perform using Fabric View:

- ▶ Managing zones and zone sets
- ▶ Managing VSANs
- ▶ Managing PortChannels
- ▶ Controlling management access with users and roles.

The menu bar at the top of the Fabric View window provides access to options, that are organized by menus. The toolbar provides icons that duplicate the most frequently used options on the File, Tools, and Help menus.

The main window has a menu bar, toolbar, message bar, status bar, and three panes:

- ▶ **VSAN/Switch pane:** Displays a tree of configured VSANs and zones on the VSANs/Zones tab and a menu tree of available configuration tasks on the Switch tab.
- ▶ **Information pane:** Displays information about whatever option is selected in the menu tree.
- ▶ **Map pane:** Displays a map of the network fabric, including switches, hosts, and storage. It also provides tabs for displaying log and event data.

6.5.10 Menu bar, toolbars and message bar

The menu bar at the top of the Fabric View window provides options for managing and troubleshooting the current fabric and for controlling the display of information on the Map pane. The menu bar provides the following menus:

- ▶ **File:** Open a new fabric, rediscover the current fabric, locate switches, set preferences, print the map, and clear or export the Map pane log.
- ▶ **Edit:** Manage zones, zonesets, and various elements on the Fabric View map.
- ▶ **View:** Change the appearance of the map (these options are duplicated on the Map pane toolbar).
- ▶ **Reports:** Display summary reports
- ▶ **Troubleshooting:** Verify and troubleshoot connectivity and configuration
- ▶ **Help:** Display on-line help topics for specific dialog boxes in the Information pane.

The Fabric View main toolbar provides buttons for accessing the most frequently used menu bar options. The Map pane toolbar provides buttons for managing the appearance of the map. The Information pane toolbar provides buttons for editing and managing the Information pane. The message bar displays the last entry displayed by the discovery process. The status bar displays the current state of the Information pane or other Fabric View components.

VSAN/Switch pane

Use the VSAN tab on the VSAN/Switch pane to manage VSANS and zones in the currently discovered fabric.

To manage zones, right-click the root folder in the VSAN tree and click **Edit Zones** from the pop-up menu. You will see the Edit Zones dialog box.

Use the **Switch** tab on the VSAN/Switch pane to display a menu tree of the options available for managing the switches in the currently discovered fabric. shown in Figure 6-20 is a example of the menu tree.

Map pane

There are three tabs on the bottom of the Map pane:

- ▶ **Map:** Displays a graphical view (Map View) of the network fabric with switches, hosts, and storage subsystems.
- ▶ **Log:** Displays messages that describe system operations, such as fabric discovery.
- ▶ **Events:** Displays information about the SNMP traps received by the management station.

When you right-click an icon, you see a pop-up menu with options that vary depending on the type of icon selected. The various options available for different objects include the following:

- ▶ Open a Device View for the selected switch.
- ▶ Open a CLI session for the selected switch.
- ▶ Copy the display name of the selected object.
- ▶ Execute a **ping** or **tracert** command for the device.
- ▶ Show or hide end devices.
- ▶ Create or delete an enclosure.
- ▶ Set the VSAN ID for an edge port (link).
- ▶ Set the trunking mode for an ISL.
- ▶ Create or add to a PortChannel for selected ISLs.

The Map pane has its own toolbar with options for saving, printing, and changing the appearance of the map. When you right-click on the map, a pop-up menu appears that provides options (duplicated on the toolbar) for changing the appearance of the map.

Using Device View and Summary View

Most tasks that you can perform with Device View can also be performed for multiple switches using the Fabric View. However, the Device View may be more convenient to use when you are working with a single switch. Also, the Device View provides more detailed information for verifying or troubleshooting device configuration than is available from the Fabric View.

When you start the Device Manager, you see the Device View, as shown in Figure 6-21.

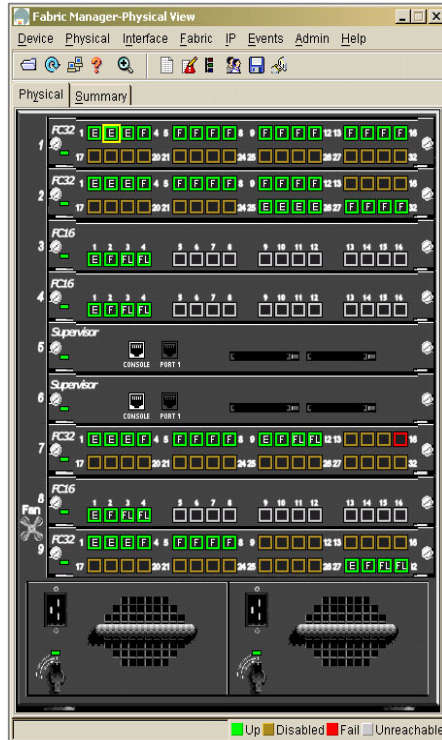


Figure 6-21 Fabric Manager - Device View

The Device View provides a graphical representation of a Cisco MDS 9000 family switch, including the installed switching modules, supervisor modules, and the status of each port within each module. You can use the Device View to perform any switch-level configuration tasks including the following:

- ▶ Manage ports, PortChannels, and trunking
- ▶ Manage SNMPv3 security access to switches
- ▶ Manage CLI security access to switches
- ▶ Manage alarms, events, and notifications
- ▶ Save and copy configuration files and software images
- ▶ View hardware configuration
- ▶ View chassis, module, and port status and statistics

Summary View provides a way of monitoring all of the ports on the switch, categorized by function. When you click the **Summary** tab on the Fabric Device Manager window, you see the Summary View, as shown in Figure 6-22.

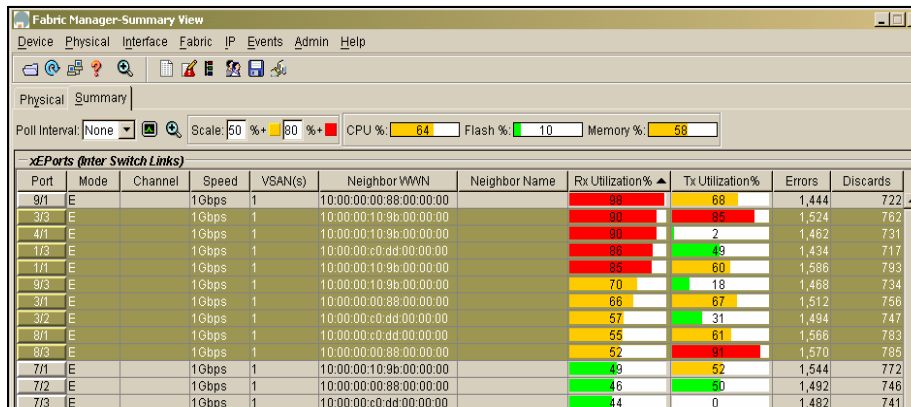


Figure 6-22 Fabric Manager - Summary View

The Summary View provides summary information about the interfaces on a single switch.

6.6 Zoning

Zoning is a mechanism of protecting resources within the SAN by grouping devices together that require common access. Zoning allows us to enable or disable certain ports within the SAN.

For example, in Figure 6-23 we show a typical scenario where a customer has a SAN with an NT server and a UNIX server. In this scenario we want to restrict the NT server to only see the FASTT700 disk subsystem and nothing else in the SAN. We also want to enable the UNIX server to see the same FASTT700 disk subsystem and the 3584 tape library. To do this, we define a zone (Zone_1) that contains the NT server and the FASTT700. We also create another zone (Zone_2) to handle allow the UNIX server to see the FASTT700 disk subsystem and the 3584 tape subsystem.

In this example we choose to do this to stop NT from acquiring all the devices attached to the SAN. In this example, if the SAN were not zoned, then there would be nothing to stop the NT server from acquiring the UNIX systems disk and tape devices.

For a more detailed overview of zoning, refer to 2.4.6, "Zoning" on page 55.

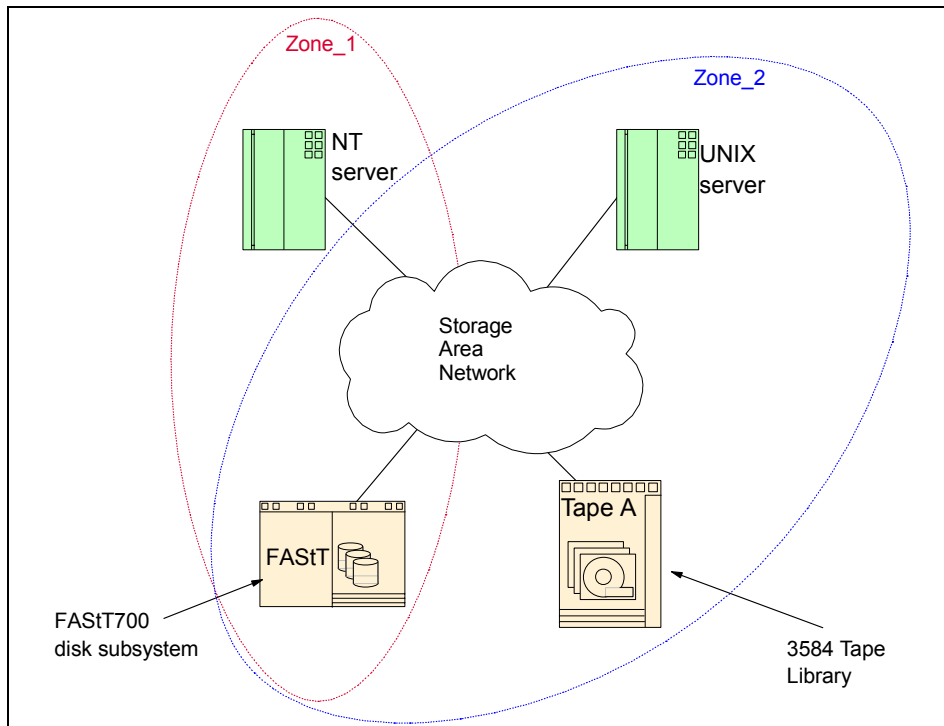


Figure 6-23 Zoning overview

6.6.1 Zone features

Zoning has the following features:

- ▶ A zone consists of multiple zone members:
 - Members in a zone can only access each other.
 - If zoning is not activated, all devices are members of the default zone.
 - If zoning is activated, any device that is not in an active zone (a zone that is part of an active zone set) is a member of the default zone.
 - Zones can vary in size.
 - Devices can belong to more than one zone.
- ▶ A zone set consists of one or more zones:
 - A zone set can be activated or deactivated as a single entity across all switches in the fabric.
 - Only one zone set can be activated at any time.
 - A zone can be a member of more than one zone set.

- ▶ Zoning can be administered from any switch in the fabric:
 - Because zoning information is distributed to all switches in the fabric, zoning changes made on one switch are available in all switches.
 - If a new switch is added to an existing fabric, zone sets are acquired by the new switch.
- ▶ Zone changes can be configured non-disruptively:
 - New zones and zone sets can be configured without interrupting traffic on unaffected ports or devices.
- ▶ The Default Zone includes all ports or WWNs that do not have a specific membership association.
 - Access between default zone members is controlled by the default zone policy.

6.6.2 Zone membership

The Cisco MDS 9000 family offer a number of methods of zoning. Zones can be set up using:

- ▶ **Port World Wide Node Name (pWWN):** This specifies the pWWN of an N_Port attached to the switch
- ▶ **Fabric pWWN:** This specifies the WWN of the fabric port (or the switch port's WWN). This is also referred to as port-based zoning.
- ▶ **Fibre Channel ID:** This specifies the Fibre Channel ID of an N_Port.

6.6.3 Configuring a zone

A zone can be configured using one of the following types to assign members:

- ▶ **pWWN:** The WWN of the N or NL port in hexadecimal format (for example, 10:00:00:23:45:67:89:ab).
- ▶ **Fabric port WWN:** The WWN of the fabric port name in hexadecimal format (for example, 10:00:00:23:45:67:89:ab).
- ▶ **FC ID:** The N port ID in 0xhhhhhh format (for example, 0xce00d1).
- ▶ **FC alias:** The alias name is in alphabetic characters (for example, *Payroll*) and denotes a port ID or WWN. The alias can also include multiple members.

6.6.4 Zone enforcement

Zoning can be enforced in two ways: soft zoning and hard zoning.

Each end device (N port or NL port) discovers other devices in the fabric by querying the name server. When a device logs in to the name server, the name server returns the list of other devices that can be accessed by the querying device. If an Nx port does not know about the FC IDs of other devices outside its zone, it cannot access those devices.

In soft zoning, zoning restrictions are applied only during interaction between the name server and the end device. If an end device somehow knows the FC ID of a device outside its zone, it can access that device.

Hard zoning is enforced by the hardware on each frame sent by an Nx port. As frames enter the switch, source-destination IDs are compared with permitted combinations to allow the frame at wirespeed.

The Cisco MDS 9000 family of switches supports both of these methods.

6.6.5 Zone sets

While zones provide access control to devices, a zone set is a group of zones to enforce access control across the whole fabric. Multiple zone sets can be created but only a single zone set can be activated at once. Zone sets contain the names of member zones.

If one zone set is currently active and another zone set is activated then the current zone set is de-activated and the new zone set comes into effect.

6.6.6 Default zone

Each member of a fabric can belong to any zone. If a member is not part of any active zone, it is considered to be part of a default zone. Therefore, if no zone set is active in the fabric, all devices are considered to be in the default zone. Even though a member can belong to multiple zones, a member that is part of the default zone cannot be part of any other zone.

The switch determines whether a port is a member of the default zone when the attached port comes up. Traffic can be permitted or denied to members of the default zone.

This information is not distributed to all switches; it must be performed for each switch. When the switch is initialized for the first time, no zones are configured and all members are considered to be part of the default zone. Members are not permitted to talk to each other.

This ensures that devices do not gain access to each other before zoning is activated.

6.7 Virtual SAN

The MDS 9000 SAN Fabric family introduces Cisco's Virtual SAN (VSAN) technology, offering the capability to overlay multiple hardware enforced virtual fabric environments within a single physical fabric infrastructure. Each VSAN contains separate (dedicated) fabric services designed for enhanced scalability, resilience, and independence among storage resource domains.

This is especially useful in segregating service operations and failover events between high availability resource domains allocated to different VSANs. Each VSAN contains its own complement of hardware-enforced zones, dedicated fabric services, and management capabilities, just as if the VSAN were configured as a separate physical fabric.

Therefore, VSANs are designed to allow more efficient SAN utilization and flexibility, because SAN resources may be allocated and shared among more users, while supporting secure segregation of traffic and retaining independent control of resource domains on a VSAN-by-VSAN basis.

VSANs offer the following features:

- ▶ **Traffic isolation:** Traffic is contained within VSAN boundaries and devices reside only in one VSAN thus ensuring absolute separation between user groups, if desired.
- ▶ **Scalability:** VSANs are overlaid on top of a single physical SAN. The ability to create several logical VSAN layers increases the scalability of the SAN.
- ▶ **Per-VSAN fabric services:** Replication of fabric services on a per-VSAN basis provides increased scalability and availability.
- ▶ **Redundancy:** Several VSANs created on the same physical SAN ensure redundancy. If one VSAN fails, redundant protection is provided by a configured backup path between the host and the switch.
- ▶ **Ease of configuration:** Users can be added, moved, or changed between VSANs without changing the physical structure of a SAN. Moving a device from one VSAN to another only requires configuration at the port level, not at a physical level.
- ▶ **Shared topology:** Multiple VSANs can share the same physical topology.
- ▶ **Same FCIDs:** The same Fibre Channel IDs (FCIDs) can be assigned to a host in another VSAN, thus increasing VSAN scalability.
- ▶ **Required protocols:** Every instance of a VSAN runs all required protocols such as FSPF, domain manager, and zoning.

- ▶ **Independence:** Fabric-related configurations in one VSAN do not affect the associated traffic in another VSAN.
- ▶ **Containment:** Events causing traffic disruptions in one VSAN are contained within that VSAN and are not propagated to other VSANs.
- ▶ **Isolation:** No communication is possible between VSANs.

Figure 6-24 represents a typical SAN environment that has a number of servers, each with multiple paths to the SAN. The SAN in this case consists of a Fibre Channel director attached to a disk and tape subsystem.

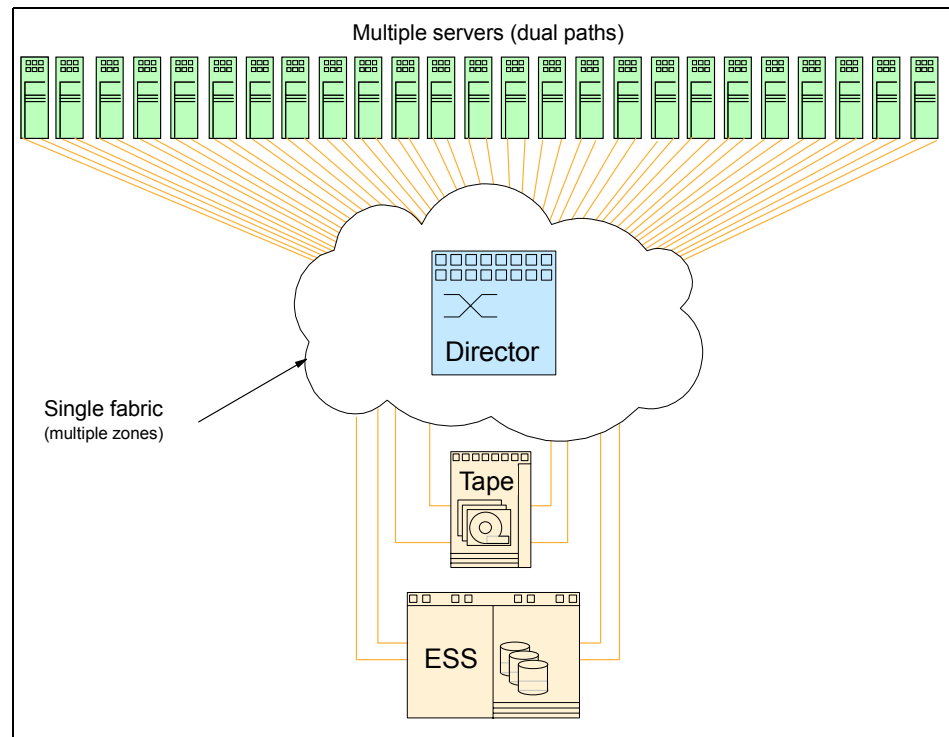


Figure 6-24 Traditional SAN

In Figure 6-25 we show how the same scenario is implemented using Cisco's Virtual SAN.

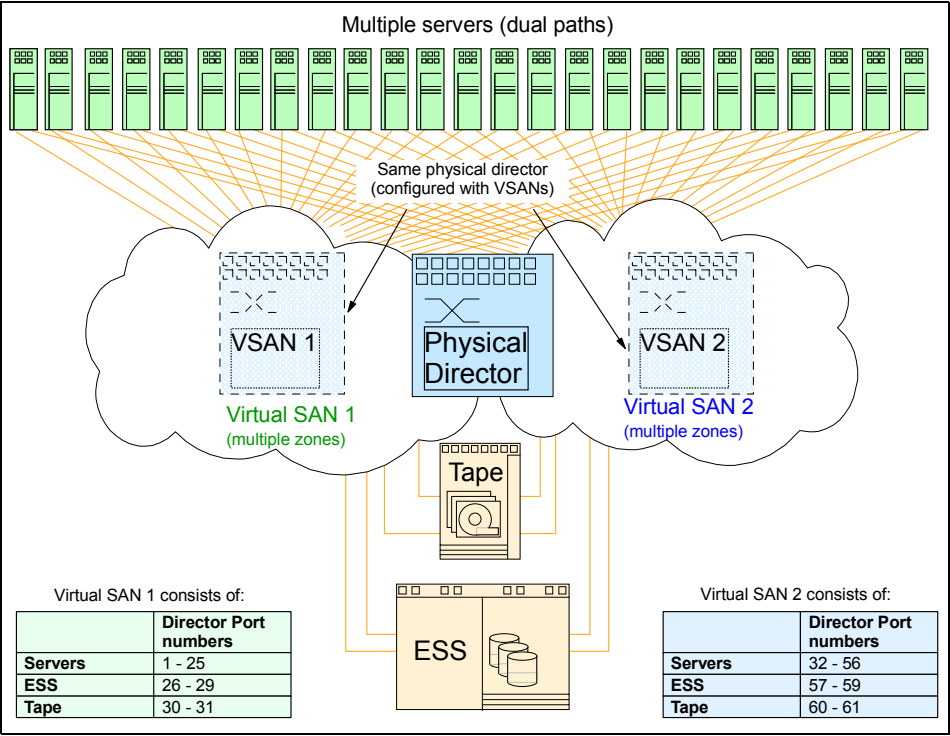


Figure 6-25 Cisco Virtual SAN

In this example, the servers are still connected to the SAN, but the SAN consists of a single MDS 9509 Director which is attached to the same disk and tape subsystems. In this case we have configured the first 31 ports in the director into a Virtual SAN (called Virtual SAN 1) and the second 31 ports into another virtual SAN (called Virtual SAN 2). The servers have a connection to each virtual SAN thereby providing a solution that consists of multiple SAN fabrics.

The virtual SANs cannot communicate with each other, they appear to be totally separate SANs. They have their own FSPF tables, domain manager and zoning requirements. A traffic disruption in one virtual SAN will have no impact on the other virtual SAN. A port cannot belong in multiple VSANs.

6.7.1 VSANs versus zones

Table 6-8 compares the main differences between VSANs and zones.

Table 6-8 VSAN versus zoning

VSANs	Zones
VSANs equal SANs with their own routing, naming, and zoning protocols.	These protocols are not always available on a per-zone basis.
VSANs can contain multiple zones.	Zones are always contained within a VSAN - they cannot span a VSAN.
VSANs limit unicast, multicast and broadcast traffic.	Zones limit unicast traffic.
Membership is defined using VSAN ID to Fx ports.	Membership is typically defined using WWN or port number (Fx).
HBAs may only belong to a single VSAN - that is the VSAN associated with the Fx port.	HBAs can belong in multiple zones.
VSANs enforce membership at each E_Port, source port and destination port.	Zones enforce membership only at the source and destination ports.
VSANs encompass the entire fabric.	Zones typically contain a subset of the fabric.

6.7.2 Default and isolated VSANs

Up to 1024 VSANs can be configured on a physical SAN. Of these, one is the default VSAN (VSAN 1) and another is an isolated VSAN (VSAN 4094). User-specified VSAN IDs range from 2 to 4093.

Default VSAN

The factory settings for switches in the Cisco MDS 9000 family have only the default VSAN 1 enabled. If you do not need more than one VSAN for a switch, use this default VSAN as the implicit parameter during configuration. If no VSANs are configured, all devices in the fabric are considered part of the default VSAN. By default, all ports are assigned to the default VSAN.

Isolated VSANs

VSAN 4094 is an isolated VSAN. All non-trunking ports are transferred to this VSAN when the VSAN to which they belong is deleted. This avoids an implicit transfer of ports to the default VSAN or to another configured VSAN. All ports in the deleted VSAN are isolated (disabled).

6.7.3 VSAN membership

Port VSAN membership on the switch is assigned on a port-by-port basis. By default each port belongs to the default VSAN. Trunking ports have an associated list of VSANs that are part of an allowed list.

6.7.4 VSAN attributes

VSANs have the following attributes:

- ▶ **VSAN ID:** The VSAN ID identifies the VSAN as the default VSAN (VSAN 1), user-defined VSANs (VSAN 2 to 4093), and the isolated VSAN (VSAN 4094).
- ▶ **State:** The administrative state of a VSAN can be configured to an active (default) or suspended state. Once VSANs are created, they may exist in various conditions or states.
 - The active state of a VSAN indicates that the VSAN is configured and enabled. By enabling a VSAN, you activate the services for that VSAN.
 - The suspended state of a VSAN indicates that the VSAN is configured but not enabled. If a port is configured in this VSAN, it is disabled. Use this state to deactivate a VSAN without losing the VSAN's configuration. All ports in a suspended VSAN are disabled. By suspending a VSAN, you can preconfigure all the VSAN parameters for the whole fabric and activate the VSAN immediately.
- ▶ **VSAN name:** This text string identifies the VSAN for management purposes. The name can be from 1 to 32 characters long and it must be unique across all VSANs. By default, the VSAN name is a concatenation of VSAN and a four-digit string representing the VSAN ID. For example, the default name for VSAN 3 is VSAN0003.
- ▶ **Load balancing attributes:** These attributes indicate the use of the source-destination ID (src-dst-id) or the originator exchange OX ID (src-dst-ox-id, the default) for load balancing path selection.

Operational state of a VSAN

A VSAN is in the operational state if the VSAN is active and at least one port is up. This state indicates that traffic can pass through this VSAN. This state cannot be configured.

Deleted VSAN

When an active VSAN is deleted, all of its attributes are removed from the running configuration.

VSAN related information is maintained by the system software:

- ▶ VSAN attributes and port membership details are maintained by VSAN manager. This feature is affected when you delete a VSAN from the configuration. When a VSAN is deleted all the ports in that VSAN are made inactive and the ports are moved to the isolated VSAN. If the same VSAN is recreated, the ports do not automatically get assigned to that VSAN. You must explicitly reconfigure the port VSAN membership.
- ▶ VSAN-based runtime (name server), zoning and configuration (static route) information is removed when the VSAN is deleted.
- ▶ Configured VSAN interface information is removed when the VSAN is deleted.

6.8 Security management

The Cisco MDS 9000 family of switches offer strict and secure switch management options through switch access security, user authentication, and role-based access.

6.8.1 Switch access security

Each switch can be accessed through the CLI or SNMP:

- ▶ **Secure switch access:** Available when you explicitly enable Secure Shell (SSH) access to the switch. SSH access provides additional controlled security by encrypting data, user IDs, and passwords. By default, Telnet access is enabled on each switch.
- ▶ **SNMP access:** SNMPv3 provides built-in security for secure user authentication and data encryption.

6.8.2 User authentication

A strategy known as authentication, authorization, and accounting (AAA) is used to verify the identity of, grant access to, and track the actions of remote users. The RADIUS protocol provides AAA solutions.

Based on the user ID and password combination provided, switches perform local authentication using a local database or remote authentication using the RADIUS server(s). A global, preshared, secret key authenticates communication between the RADIUS client and server. This secret key can be configured for all RADIUS servers or for only a specific RADIUS server. This kind of authentication provides a central configuration management capability.

6.8.3 Role-based access

Role-based access assigns roles or groups to users and limits access to the switch. Access is assigned based on the permission level associated with each user ID. Your administrator can provide complete access to each user or restrict access to specific read and write levels for each command.

SNMP and CLI access rights are organized by roles. Each role is similar to a group. Each group of users has a specific role, and the access for that group can be enabled or disabled.

6.8.4 User authentication

Authentication is the process of verifying the identity of the person managing the switch. This identity verification is based on the user ID and password combination provided by the person trying to manage the switch. Cisco MDS 9000 family switches allow you to perform local authentication (using the lookup database) or remote authentication (using one or more RADIUS servers).

For each management path (console or Telnet and SSH), you can enable only one of three options — local, RADIUS, or none. The option can be different for each path.

Local authentication

The system maintains the user name and password locally and stores the password information in encrypted form. You are authenticated based on the locally stored information, as shown in Figure 6-26.

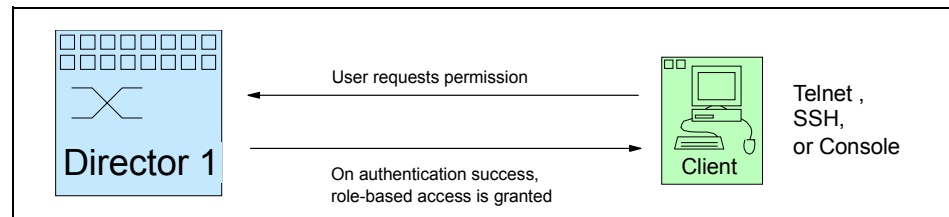


Figure 6-26 Security with local authentication

RADIUS authentication

Cisco MDS 9000 family switches can provide remote authentication through RADIUS servers. You can also configure multiple RADIUS servers, and each server is tried in the order specified.

RADIUS protocols support one-time password (OTP) schemes that all switches can make use of for authentication purposes.

The use of RADIUS servers in the authentication process is shown in Figure 6-27.

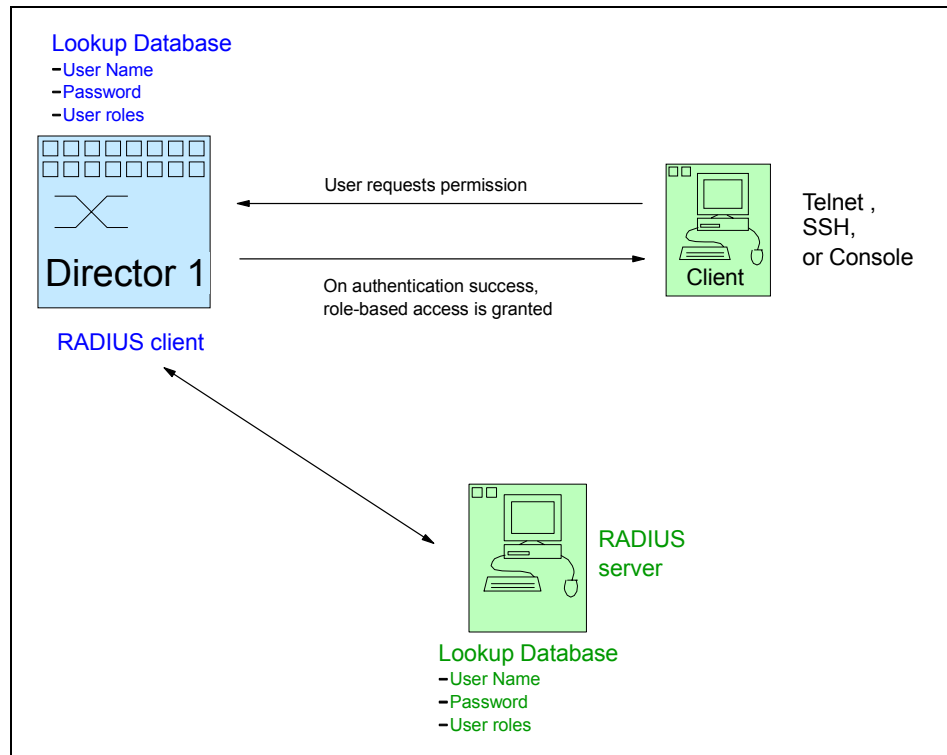


Figure 6-27 Security with RADIUS server

6.8.5 Role-based authorization

By default, two roles exist in all switches:

- ▶ **Network operator (network-operator):** Has permission to view the configuration only. The operator cannot make any configuration changes.
- ▶ **Network administrator (network-admin):** Has permission to execute all commands and make configuration changes. The administrator can also create and customize up to 64 additional roles.

The two default roles cannot be changed or deleted. Vendor-specific attributes (VSAs) contain the user profile information used by the switch. To use this option, configure the VSAs on the RADIUS servers.

6.8.6 Accounting

Accounting refers to the log that is kept for each management session in a switch. This information may be used to generate reports for troubleshooting purposes and user accountability. Accounting can be implemented locally and remotely (using RADIUS).

6.8.7 Standards

These are the FC standards that Cisco supports.

- ▶ FC-PH, Revision 4.3
- ▶ FC-PH-2, Revision 7.4
- ▶ FC-PH-3, Revision 9.4
- ▶ FC-GS-2, Revision 5.3
- ▶ FC-GS-3, Revision 7.01
- ▶ FC-FLA, Revision 2.7
- ▶ FC-FG, Revision 3.5
- ▶ FC-SW-2, Revision 5.3
- ▶ FC-AL, Revision 4.5
- ▶ FC-AL-2, Revision 7.0
- ▶ FC-PLDA, Revision 2.1
- ▶ FC-VI, Revision 1.61
- ▶ FCP, Revision 12
- ▶ FCP-2, Revision 7a
- ▶ FC-SB-2, Revision 2.1
- ▶ FC-BB, Revision 4.7
- ▶ FC-FS, Revision 1.7
- ▶ FC-PI, Revision 13
- ▶ FC-MI, Revision 1.99
- ▶ FC-Tape, Revision 1.17

6.9 Switch interoperability mode

Switch interoperability mode enables multiple vendors' products to connect to each other. This mode is required because not all vendors' switches follow the standards in the same manner, so this mode provides a common implementation that all vendors' switches can apply.

Switch interoperability mode will specifically turn off a number of advanced or proprietary features so it is worth understanding what these effects may be before proceeding down this path.

Table 6-9 lists the changes required should interoperability mode be enabled on a Cisco MDS 9000 family switch or director.

Table 6-9 Interoperability mode changes

Switch feature	Changes if interoperability mode is enabled
Domain IDs	Some vendors cannot use the full range of 239 domains within a fabric. Domain IDs are restricted to the range 97-127 to accommodate all vendors implementations.
Timers	All Fibre Channel timers must be the same on all switches as these values are exchanged by E_Ports when establishing an ISL. The Time-Out Value timers are described in the following rows.
F_S_TOV	Verify that the Fabric Stability Time-Out Value timers match exactly.
D_S_TOV	Verify that the Distributed Services Time-Out Value timers match exactly.
E_D_TOV	Verify that the Error Detect Time-Out Value timers match exactly.
R_A_TOV	Verify that the Resource Allocation Time-Out Value timers match exactly.
Trunking	Trunking is not supported between two different vendors switches. This feature may be disabled on a per port basis.
Default Zone	The default zone behavior of permit (all nodes can see other nodes) or deny (all nodes are isolated when not explicitly placed in a zone), may change.
Zoning attributes	Zones may be limited to the WWPN and other proprietary zoning methods (physical port number), may be eliminated.
Zone propagation	Some vendors do not pass the full zone configuration to other switches, only the active zoneset gets passed.
VSAN	Will only affect the specified VSAN.
TE_Ports and PortChannels	TE_Ports and PortChannels only apply when connecting from one MDS 9000 to another. Only E_Ports can be used to connect to non-MDS switches. TE_Ports and PortChannels can be used to connect to other MDS 9000 switches when Interoperability Mode is enabled.
Domain reconfiguration disruptive	This can require the entire switch to be restarted when changing Domain IDs.

Switch feature	Changes if interoperability mode is enabled
Domain configuration nondisruptive	This will only impact the affected VSAN. Only the domain manager for the affected VSAN is restarted but other VSANs will be unaffected.
Name Server	Need to verify that all vendors have the correct values in their respective Name Server tables.

Interoperability mode in the Cisco MDS 9000 family can be enabled non-disruptively, but the default is to have this mode turned off.

It is still important to check with the OEM vendors involved as to the specific steps that must be taken.

6.10 Troubleshooting

The Cisco MDS 9000 family of switches and directors provide a comprehensive range of diagnostic and troubleshooting features. The components within the chassis have status LEDs that can be used to determine the status of the various components including fans, power supplies, switching modules and supervisor modules. As well as the LEDs, the MDS 9000 family provide extensive software tools that assist with the troubleshooting process some of these are discussed in the following section. For more detailed information, refer to:

- ▶ *Cisco MDS 9000 Family Fabric Manager User Guide, 78-14895-01*
- ▶ *Cisco MDS Family Configuration Guide, 78-14893-01*

6.10.1 Troubleshooting with Fabric Manager

In this section we describe some of the tools provided by the Fabric View and Device View that can be used to verify and troubleshoot fabric connectivity and switch configuration issues.

Switch health option

Allows you to determine the status of the components of a specific switch. The Switch Health Analysis window will display any problems affecting the selected switches.

Analyzing end-to-end connectivity

This option can be used to check connectivity among devices within the SAN fabric. The tool checks to see that every pair of end devices in an active zone can talk to each other.

This is achieved using the **FC Ping** and **Traceroute** commands which have been modified to handle Fibre Channel networks.

The End-to-End Connectivity Analysis window displays the selected end points of the switch and to which end point each is attached, and the source and target ports used to connect it.

The output shows all the requests which have failed. Descriptions possible are:

- ▶ **Ignoring empty zone:** No requests are issued for this zone.
- ▶ **Ignoring zone with single member:** No requests are issued for this zone.
- ▶ **Source/Target are unknown:** No name server entries exist for the ports or the ports have not been discovered during discovery.
- ▶ **Both devices are on the same switch:** The devices are not redundantly connected.
- ▶ **No paths exist:** (self-explanatory).
- ▶ **Only one unique path exists:** (self-explanatory).
- ▶ **VSAN does not have an active zone set:** (self-explanatory).
- ▶ **Average time (in microseconds):** The latency value was more than the threshold supplied.

Analyze switch fabric configuration

This option enables you to analyze the configuration of a switch by comparing the current configuration to a specific switch or to a policy file. You can save a switch configuration to a file and then compare all switches against the configuration in this file.

Analyzing the results of a zone merge

You can use the Zone Merge option on the Fabric View Troubleshooting menu to determine if two connected switches have compatible zone configurations.

Other troubleshooting tools

Other troubleshooting tools include:

- ▶ **Traceroute command:** Used to verify the connectivity between two end devices that are currently selected on the Map pane.
- ▶ **Device Manager:** Launch the Fabric Device Manager for the switch selected on the Map pane.
- ▶ **Command Line Interface:** Open a Telnet or SSH session for the switch selected on the Map pane.

6.10.2 Monitoring network traffic using SPAN

The Cisco MDS 9000 family provides a feature called the switch port analyzer (SPAN). As mentioned in 6.1.3, “Supported port-types” on page 188, the SPAN or SD_Ports allow us to monitor network traffic through the Fibre Channel interface.

Traffic through any Fibre Channel interface can be replicated to a special port called the SPAN destination port. Any Fibre Channel port in a switch can be configured as an SD_Port. Once an interface is in SD_Port mode, it cannot be used for normal data traffic. You can attach a Fibre Channel analyzer to the SD_Port to monitor SPAN traffic.

SD_Ports do not receive frames, they only transmit a copy of the SPAN source traffic. The SPAN feature is non-intrusive and does not affect switching of network traffic for any SPAN source port.

Illustrated in Figure 6-28 is an overview of the SPAN port.

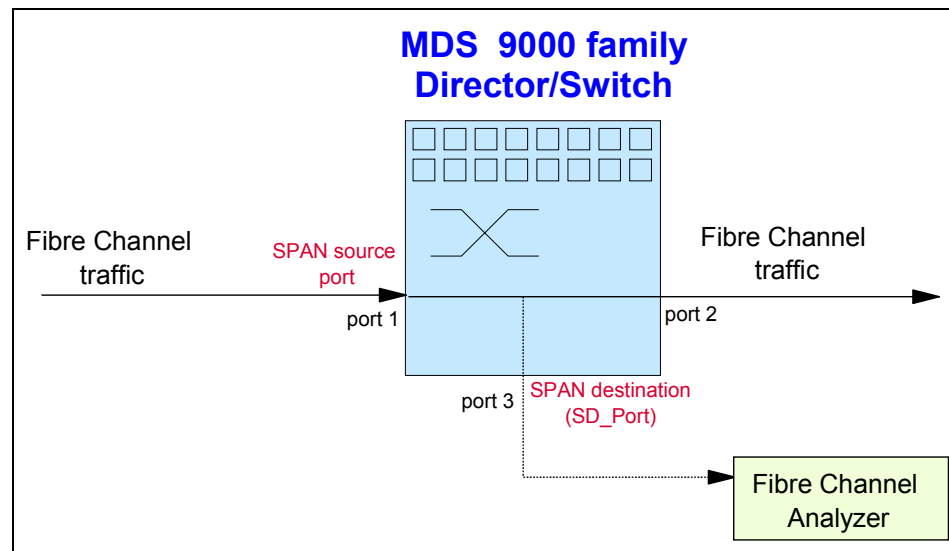


Figure 6-28 SPAN destination ports

SPAN sources

A SPAN source is the interface from which traffic can be monitored. You can also specify VSAN as a SPAN source, in which case, all supported interfaces in the specified VSAN are included as SPAN sources. You can choose the SPAN traffic in the ingress direction, the egress direction, or both directions, for any source interface.

- **Ingress source (rx):** Traffic entering the switch fabric through this source is spanned or copied to the SD_Port, as shown in Figure 6-29.

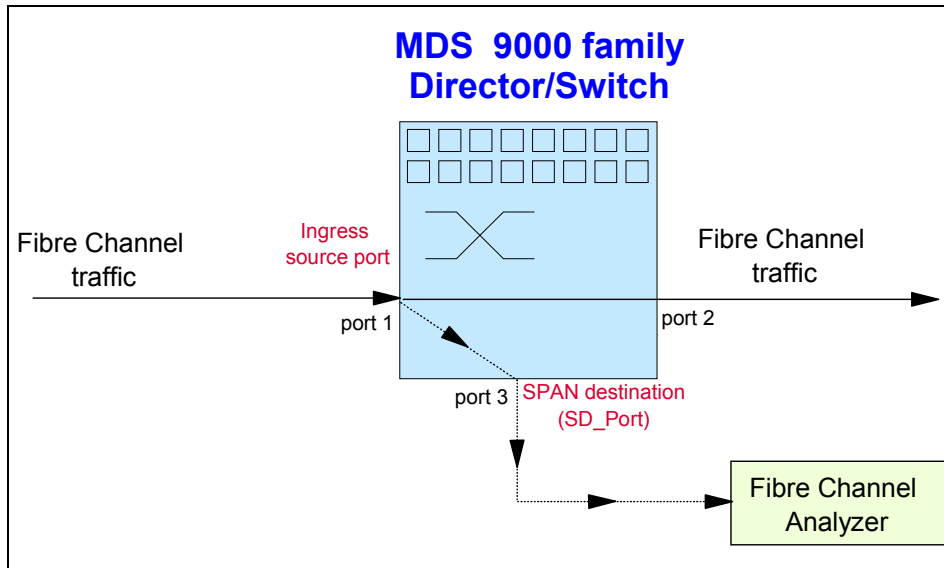


Figure 6-29 SD_Port for incoming traffic (ingress direction)

- **Egress source (tx):** Traffic exiting the switch fabric through this source interface is spanned or copied to the SD_Port, as shown in Figure 6-30.

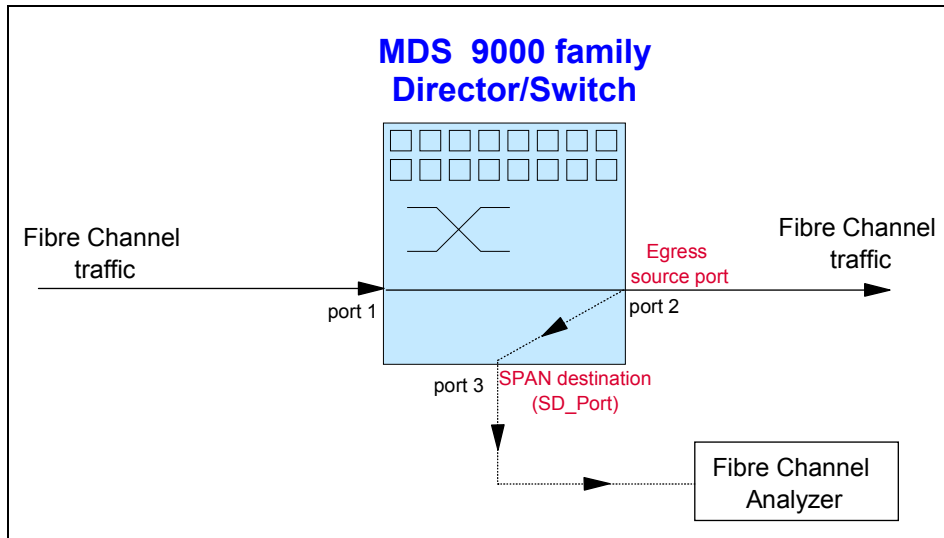


Figure 6-30 SD_Port for outgoing traffic (egress direction)

Allowed source interface types

The SPAN feature is available for the following interface types:

- ▶ Physical ports:
 - F_Ports, FL_Ports, TE_Ports, E_Ports, and TL_Ports.
- ▶ Interface sup-fc0 (traffic to and from the supervisor):
 - The Fibre Channel traffic from the supervisor module to the switch fabric, through the sup-fc0 interface, is called ingress traffic. It is spanned when sup-fc0 is chosen as an ingress source port.
 - The Fibre Channel traffic from the switch fabric to the supervisor module, through the sup-fc0 interface, is called egress traffic. It is spanned when sup-fc0 is chosen as an egress source port.
- ▶ PortChannels:
 - All ports in the PortChannel are included and spanned as sources.
 - You cannot specify individual ports in a PortChannel as SPAN sources. Previously-configured SPAN-specific interface information is discarded.

VSAN as a SPAN source

When a VSAN as a source is specified, then all physical ports and PortChannels in that VSAN are included as SPAN sources. A TE_Port is included only when the port VSAN of the TE_Port matches the source VSAN. A TE_Port is excluded even if the configured allowed VSAN list may have the source VSAN, but the port VSAN is different.

Guidelines for configuring VSANs as a source

The following guidelines apply when configuring VSANs as a source:

- ▶ Traffic on all interfaces included in a source VSAN is spanned only in the ingress direction.
- ▶ When a VSAN is specified as a source, you will not be able to perform interface-level configuration on the interfaces that are included in the VSAN. Previously-configured SPAN-specific interface information is discarded.
- ▶ If an interface in a VSAN is configured as a SPAN source, you will not be able to configure that VSAN as a source. You must first remove the existing SPAN configurations on such interfaces before configuring VSAN as a source.
- ▶ Interfaces are only included as sources when the port VSAN matches the source VSAN.

SPAN sessions

Each SPAN session represents an association of one destination with a set of source(s) along with various other parameters that you specify to monitor the network traffic. One destination can be used by one or more SPAN sessions. You can configure up to 16 SPAN sessions in a switch. Each session can have several source ports and one destination port.

To activate a SPAN session, at least one source and the SD_Port must be up and functioning. Otherwise, traffic will not be directed to the SD_Port.

To temporarily deactivate (suspend) a SPAN session use the **suspend** command in the SPAN submode. The traffic monitoring is stopped during this time. You can reactivate the SPAN session using the **no suspend** command.

Specifying filters

You can perform VSAN-based filtering to selectively monitor network traffic on specified VSANs. You can apply this VSAN filter to the selected source or to all sources in a session. Only traffic in the selected VSANs is spanned when you configure VSAN filters. You can specify two types of VSAN filters:

- ▶ **Interface level filters:** You can apply VSAN filters for a specified TE_Port or trunking PortChannel to filter traffic using one of three options — the ingress direction, the egress direction, or both directions.
- ▶ **Session filters:** This option filters all sources in the specified session. These filters are bidirectional and apply to all sources configured in the session.

Guidelines for specifying filters

The following guidelines apply to SPAN filters:

- ▶ Specify filters in either the ingress direction, or in the egress direction, or in both directions.
- ▶ PortChannel filters are applied to all ports in the PortChannel.
- ▶ If no filters are specified, the traffic from all active VSANs for that interface is spanned.
- ▶ The effective filter on a port is the intersection (filters common to both) of interface filters and session filters.
- ▶ While you can specify any arbitrary VSAN filters in an interface, traffic can only be monitored on the port VSAN or on allowed-active VSANs in that interface.
- ▶ When you configure VSAN as a source, that VSAN is implicitly applied as an interface filter to all sources included in the specified VSAN.

SD_Port characteristics

An SD_Port has the following characteristics:

- ▶ It ignores buffer-to-buffer credits.
- ▶ It allows data traffic only in the egress (tx) direction.
- ▶ It does not require a device or an analyzer to be physically connected.
- ▶ It supports only 1 Gb/s or 2 Gb/s speeds. The auto speed option is not allowed.
- ▶ Multiple sessions can share the same destination ports.
- ▶ If the SD_Port is shut down, all shared sessions stop generating SPAN traffic.
- ▶ The port mode cannot be changed if it is being used for a SPAN session.
- ▶ The outgoing frames can be encapsulated in EISL format.
- ▶ The SD_Port does not have a port VSAN.

The following guidelines apply for a SPAN configuration:

- ▶ You can configure up to 16 SPAN sessions with multiple ingress (rx) sources.
- ▶ You can configure a maximum of three SPAN sessions with one egress (tx) port.
- ▶ In a 32 port switching module, you must configure the same session in all four ports in one port group. If you wish, you can also configure only two or three ports in this unit
- ▶ SPAN frames are dropped if the sum of the bandwidth of the sources exceeds the speed of the destination port.
- ▶ Frames dropped by a source port are not spanned.

6.10.3 Monitoring traffic using Fibre Channel analyzers

You can use SPAN to monitor traffic on an interface without any disruption to traffic. This feature is very useful in troubleshooting scenarios when traffic disruption can change the problem environment, and it can also make it difficult to reproduce the problem.

Without SPAN

You can monitor traffic, as shown in Figure 6-31, using interface 1 in a Cisco MDS 9000 family switch that is connected to another device on the SAN. You would need to physically connect a Fibre Channel analyzer between the switch and the storage device to analyze the traffic through interface 1.

This type of connection has the following limitations:

- ▶ It requires you to insert the Fibre Channel analyzer between the devices, which disrupts traffic.
- ▶ The Fibre Channel analyzer captures data only on the rx links in both port A and port B. Port A captures traffic exiting interface 1 and port B captures ingress traffic into interface 1.

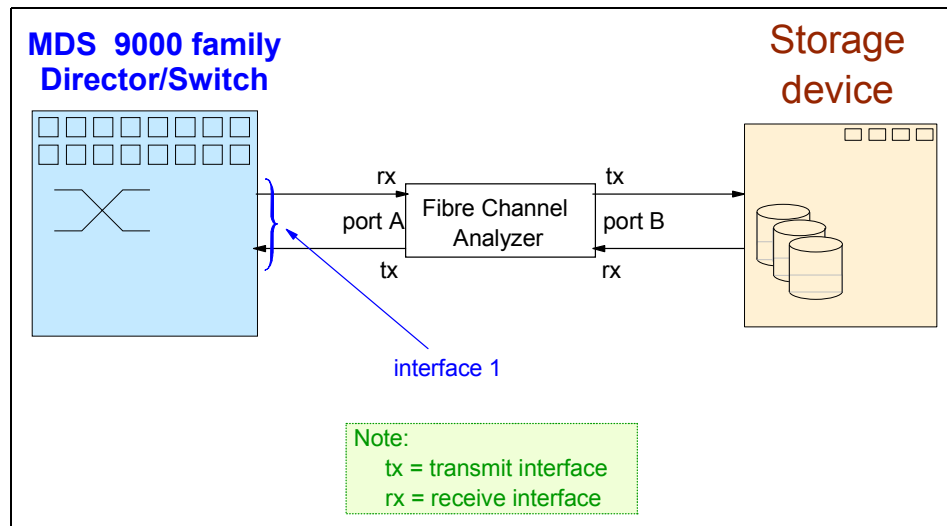


Figure 6-31 Fibre Channel analyzer without SPAN

Using SPAN

Using SPAN, you can capture the same traffic scenario shown in Figure 6-31 on page 250 without any traffic disruption. The Fibre Channel analyzer uses the ingress (rx) link at port A to capture all the frames going out from port 1. It uses the ingress link at port B, to capture all the frames coming in on port 1. This is illustrated in the diagram in Figure 6-32.

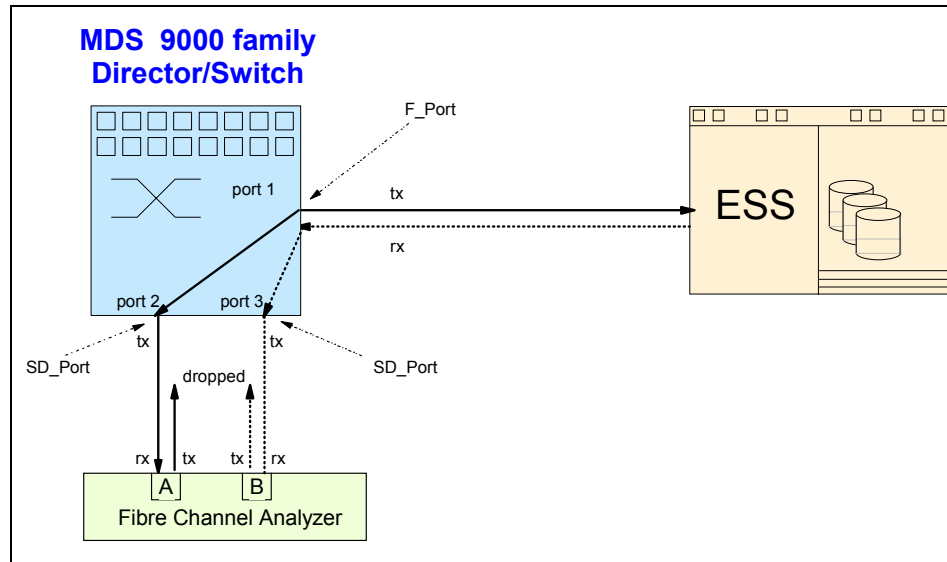


Figure 6-32 Fibre Channel analyzer using SPAN

Using a single SD_Port to monitor traffic

You do not need to use two SD_Ports to monitor bidirectional traffic on any interface, as shown in Figure 6-32. You can use one SD_Port and one FC analyzer port by monitoring traffic on the interface at the same SD_Port (Port 2).

Figure 6-33 shows a SPAN setup where one session with a destination of Port 2 and a source interface of Port 1 is used to capture traffic in both ingress and egress directions. This setup is more advantageous and cost-effective than the setup shown in Figure 6-32 — it uses one SD_Port and one port on the analyzer, instead of using a full, two-port analyzer.

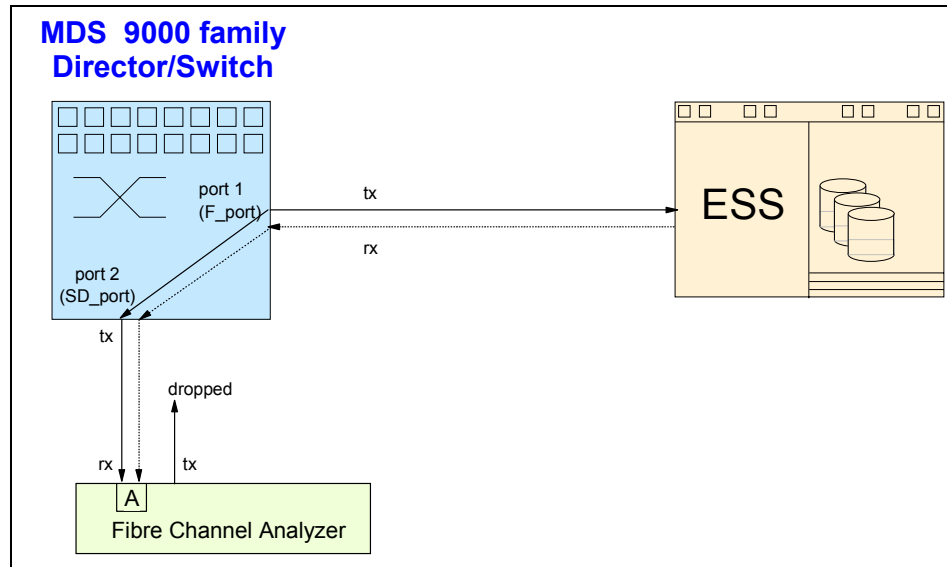


Figure 6-33 Using a single SD_Port to monitor traffic

6.11 Interoperability matrix

In Table 6-10 we provide an outline of the supported devices, including server platforms, disk systems, tape systems, and operating system levels for the Cisco MDS 9509 Multilayer Director (IBM 2062-D07).

Table 6-10 Cisco MDS 9509 Multilayer Director interoperability matrix

Cisco MDS 9509 Multilayer Director (IBM 2062-D07) - Fibre Channel support		
Storage Systems	Description	Platform/Operating Systems
Disk Systems	ESS (IBM 2105-800, F10,F20) FASiT700 (IBM 1742) FASiT500 (IBM 3552) FASiT200 (IBM 3542)	IBM pSeries (AIX 4.3.3 or later) IBM xSeries (running Linux, Netware ¹ and Windows NT/2000) Intel-based servers (running Linux, Netware ¹ and Windows NT/2000) SUN servers (running Solaris 2.6,7,8) HP (running HP-UX 11.0 & 11.i)
Tape Systems	Enterprise Tape System (IBM 3590 and 3494) Ultrascable Tape Library (IBM 3584) Scalable Tape Library (IBM 3583)	IBM pSeries (AIX 4.3.3 or 5.1) IBM xSeries (running Windows NT/2000) Intel-based servers (running Windows NT/2000) SUN servers (running Solaris 2.6, 7, 8) HP (running HP-UX 11.0 & 11.i)

Note 1: Netware support is expected in 3Q03. Refer to online interoperability matrix for the latest information.

As this list is being constantly updated, refer to the following Web site for the latest information:

<http://www.storage.ibm.com/ibmsan/products/cisco/index.html>

Selecting the Interoperability Matrix will display a complete list of servers, disk and tape systems that are supported with the MDS 9509 Multilayer Director. This list also contains supported operating system versions and links to other Web sites that document the required HBA levels.

The Cisco MDS 9216 Multilayer Fabric Switch (IBM 2062-D01) is shown in Table 6-11.

Table 6-11 Cisco MDS 9216 Multilayer Switch interoperability matrix

Cisco MDS 9216 Multilayer Switch (IBM 2062-D01) - Fibre Channel support		
Storage Systems	Description	Platform/Operating Systems
Disk Systems	ESS (IBM 2105-800, F10,F20) FAStT700 (IBM 1742) FAStT500 (IBM 3552) FAStT200 (IBM 3542)	IBM pSeries (AIX 4.3.3 or later) IBM xSeries (running Linux, and Windows NT/2000) Intel-based servers (running Linux, and Windows NT/2000) SUN servers (running Solaris 2.6,7,8) HP (running HP-UX 11.0 & 11.i)
Tape Systems	Enterprise Tape System (IBM 3590 and 3494) Ultrascalable Tape Library (IBM 3584) Scalable Tape Library (IBM 3583)	IBM pSeries (AIX 4.3.3 or 5.1) IBM xSeries (running Windows NT/2000) Intel-based servers (running Windows NT/2000) SUN servers (running Solaris 2.6, 7, 8) HP (running HP-UX 11.0 & 11.i)

As this list is being constantly updated, refer to the following Web site for the latest information:

<http://www.storage.ibm.com/ibmsan/products/cisco/index.html>

Selecting the Interoperability Matrix will display a complete list of servers, disk, and tape systems that are supported with the MDS 9216 Multilayer Switch. This list also contains supported operating system versions and links to other Web sites that document the required HBA levels.



General SAN troubleshooting tips

In this chapter, we discuss general methods to determine the root cause of faults in a SAN. Networks are usually described as *complex* — which is true — but they are not necessarily *complicated*. However, complexity means that there is the presence of a structure. We have to utilize this structure, and understand it, in order to troubleshoot the network.

This chapter is not meant to replace the numerous maintenance and service manuals which come with the SAN products. Our aim is to help identify the type of data to look at if trouble occurs. You can navigate to the various product manuals from this Web site:

<http://www.storage.ibm.com/ibmsan/index.html>

7.1 Overview

SANs embrace servers, fabric, and storage devices, although SAN administration sometimes does not cover the whole range of related products and services. Depending on the size and structure of a company's IT operation, SAN administration may be responsible for switches and storage only. Servers may be administrated by a different group, and are often managed by a group focusing on applications. Such an administrative structure can introduce delays while searching for problem root causes.

For example, a SAN problem may appear as an application problem and will be treated as such until the application is ruled out as the cause. At that stage, it will probably be handed over to the SAN group responsible.

7.2 Reporting failure

The SAN administration group may receive information from an end user as basic as: "I cannot access my storage" — for a variety of possible failures. As the underlying Fibre Channel is transparent to the end user, most times a problem in the SAN will manifest itself as a SCSI storage failure to that end user.

Depending on the type and the reporting of such a fault — it may be an automated event notification from the management station, or a user's complaint — the SAN administrator should start and check event and alarm logs in server, switches, and storage devices for failure entries.

Switches provide the internal visibility medium to the fabric, as well as to both the server and storage side. Compared to the average amount and type of information you may receive from a server that cannot access its storage, you may find the information collected in the switch to be superior to that reported on the server.

In Figure 7-1 we show an example of a fabric that we will use to give some troubleshooting tips.

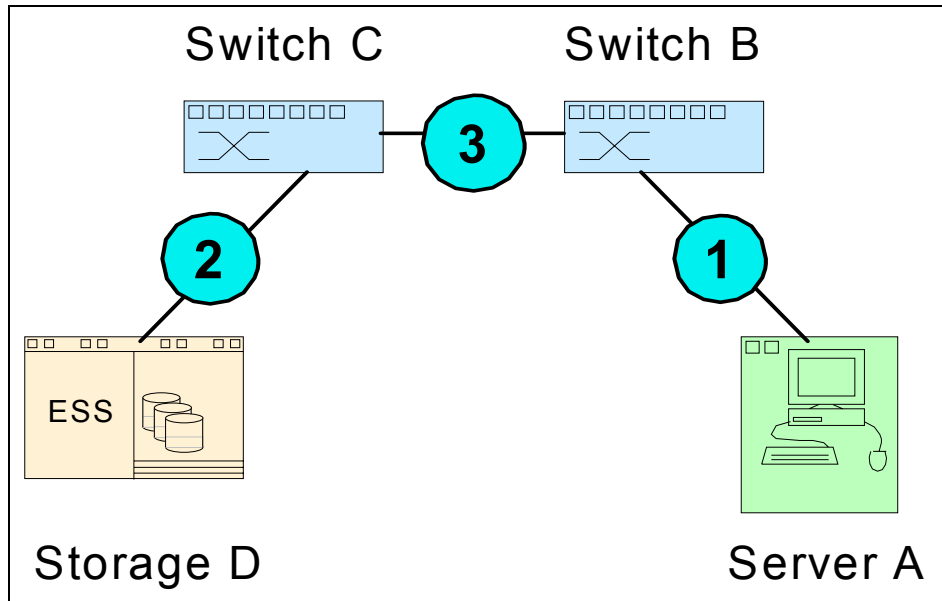


Figure 7-1 A simple fabric

- ▶ A fiber optic link problem between server and fabric (at point 1) will give you:
 - A link failure indication at server A
 - A link failure indication at switch B
 - A SCSI failure indication at server A
- ▶ A fiber optic link problem between storage and fabric (at point 2) will give you:
 - A link failure indication at storage D
 - A link failure indication at switch C
 - A SCSI failure indication at server A
- ▶ An ISL fiber optic link problem between the switches (at point 3) will give you:
 - A link failure indication at switch B
 - A link failure indication at switch C
 - A SCSI failure indication at server A
- ▶ An internal SCSI problem in server A or storage D may give you:
 - A SCSI failure indication at server A

As you can see, the internal SCSI failure is the least informative, if there is no further data provided that you can check it against. Rather simple problems may be determined by interpreting just the data of the server for instance, but most likely you'll need to see more than just one side of a connection.

Looking at Figure 7-1 on page 257, you will see there is at least one switch in each link segment from server to storage. Switches are the devices where you will get the most information. The switches are the focal point of SAN management software and switches, as provider of the fabric service, hold most of the vital SAN data: so, switches are the obvious place to look at to perform problem determination (PD) and problem source identification (PSI).

7.3 Where to look for failures

The examples in the following sections have been arbitrarily picked to highlight a particular item. They are not taken from one sample problem.

7.3.1 Connectivity problems

Let's assume you are troubleshooting a connectivity problem between server and storage (it does not matter if it is a new implementation or a failure during normal operation). Depending on the reported fault, you might omit some of the tasks from the list and utilize a different order to us for troubleshooting:

► Server

- Check the event-log and alarm-log (AIX is the example shown in Figure 7-2).

```

LABEL:      FCS_ERR4
IDENTIFIER:  B8113DD1

Date/Time:   Thu Dec 19 17:03:01 CST
Sequence Number: 724
Machine Id:  0000000000001
...
Type:        TEMP
Resource Name: fcs1
Resource Class: adapter
VPD:
             Part Number.....09P5079
             EC Level.....A
             ...

Description
LINK ERROR
```

Figure 7-2 AIX error log

- Check the status of the system and of the HBAs (pSeries is the example shown in Figure 7-3).

```
lsdev -C | grep fcs

fcs0 Defined 14-08 FC Adapter
fcs1 Defined 21-08 FC Adapter
fcs2 Available 2A-08 FC Adapter
```

Figure 7-3 pSeries HBA status

- If multi-path software is installed, check the virtual paths and adapters (SDD is the example shown in Figure 7-4).

```
datapath query device

DEV#: 0 DEVICE NAME: vpath37 TYPE: 2105F20 SERIAL: 00000001
POLICY: Optimized
=====
Path#      Adapter/Hard Disk      State  Mode  Select  Errors
0          fscsi5/hdisk41        DEAD  OFFLINE  129547   100
1          fscsi5/hdisk105       DEAD  OFFLINE  129626   100
2          fscsi5/hdisk169       DEAD  OFFLINE  129659   100
3          fscsi5/hdisk233       DEAD  OFFLINE  128803   100
...
```

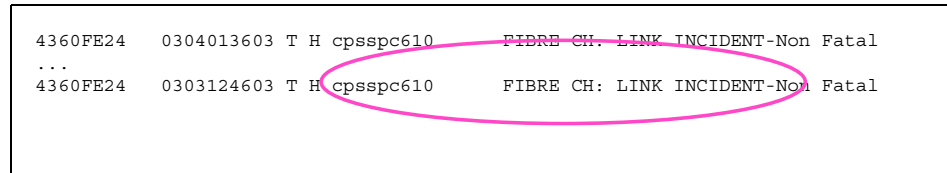
Figure 7-4 SDD information

► Switches

- Check the event and alarm log
- Look at the status of the ports to the Fibre Channel device in question and of the ISL ports
- Look at the error counter of the ports to the Fibre Channel device in question and of the ISL ports
- Check the status of the switch and the integrity of the fabric
- Check the name server table of the switches to check if the Fibre Channel devices are logged in to the fabric properly
- Check the integrity of the zoning to make sure that devices have not been isolated and the zoning config that is active, is the one that you expect to be active.

► Storage

- Check the event and alarm log of the Fibre Channel storage (ESS is the example shown in Figure 7-5).



```
4360FE24 0304013603 T H cpsspc610 FIBRE CH: LINK INCIDENT-Non Fatal
...
4360FE24 0303124603 T H cpsspc610 FIBRE CH: LINK INCIDENT-Non Fatal
```

Figure 7-5 ESS error log

- Check the status of the Fibre Channel ports.
- Check the volume assignment at the storage device.

Error logs may hold entries from the past few months so make sure you concentrate on the actual fault and do not get lost or side-tracked in all the information. Try to link the errors from the different devices together to follow the trace of the problem through the SAN; information about the components, their location and adjacent devices and/or ports are provided by the logs.

While checking the status of server, switches and storage, note down the OS version and maintenance level, driver level and firmware versions. Don't forget to *save* all this information! You may not find the problem by yourself and you may need to hand it over to other support personnel. In the meantime, some information may already be dropping off the logs or be overwritten by new entries. It is important to see data from each device along the chain from server to storage taken at the same time to get correlated failure information. It also can be of use to have a trace or information from a time when an error was not present, so that it can be compared to the error situation.

Note on preventive action: Make sure all SAN devices run on the same date/time-base, so during PD/PSI, cause and effect of the failure can be clearly correlated.

7.3.2 Performance problems

Different from a pure (and obvious) connection problems (for example a disconnected fiber optic cable) are performance problems. Some possible causes put down to performance may be:

- End-user applications needing more bandwidth than expected
- Lack of available bandwidth due to link degradation

A degraded link or path is a kind of connectivity failure which takes away some of the available bandwidth, such as a partial link out of a trunk, or a complete ISL/trunk, which loads more traffic over the remaining ISLs.

Bit errors and performance

The bit stream of all data traffic in Fibre Channel is validated by a cyclic redundancy check (CRC). The application receiving Fibre Channel frames can trust implicitly that no content has changed due to bad link quality, as the error frames are either corrected or deleted. The Fibre Channel standard limits the maximum bit error rate to “1exp-12”, which means: there must not be more than 1 bit error out of 1exp12 bits in total. According to that value, with a bandwidth of 1Gb/s of your fiber optic link, you may expect no more than 1 bit error each 1000 sec, or about 17 minutes.

This formula is shown in Figure 7-6.

$$t_{err} = \frac{1}{\text{error rate} \times \text{bandwidth}} = \frac{1}{1 \exp -12 \left(\frac{b}{s} \right) \times 1 \exp 9 b} = 1000s$$

Figure 7-6 Bit error formula

Bit errors may happen and may affect the data frames. These data frames will be re-transmitted at the request of the upper-layer protocols. If the link suffers a lot of bit errors, you may experience a slight performance loss.

These bit errors may affect the Receiver Ready (R_RDY) too. A R_RDY is never repeated, so the buffer credit is one BB_Credit short until the link is reset. If your server only had two BB_Credits at the beginning, the server will lose 50% of its credits and may suffer up to 50% performance loss (this will not happen in arbitrated loops where BB_Credits are increased at the start of each loop tenancy).

CRC: Cyclic redundancy check is an single-error correction, and multiple-error detection method. As a 32-bit value, CRC is part of the Fibre Channel frame.

Performance problem troubleshooting

When looking at performance problems, you are aiming to solve a different type of connectivity problem. So, the data sources to look at are the same as before:

► Server

- Check the event and alarm log.
- If multi-path software is installed, check the virtual paths and adapters.

► Switches

- Check the event and alarm log.
- Monitor performance on the ports to the Fibre Channel device in question and on the ISL ports. If you don't have comparable values preserved to check against, you may need to take two snapshots of the values in an interval of a couple of minutes to see how these values change over time.
- Check the status of the switch and the integrity of the fabric.
- Check the integrity of the zoning.

► Storage

- Check the event and alarm log.
- Check the status of the Fibre Channel ports.

All this information should give you an idea if there is a bottleneck by design in your SAN, or if a failing component caused that bottleneck. In the former case, perhaps an accepted ISL-oversubscription (refer to 8.7, “Definitions” on page 280) was too high and needs to be adjusted. You will want to collect the performance data for a review and potentially add more paths.

Note on preventive action: Check and save performance profiles at particular times and days to help distinguish performance problems related to design from other ones, such as those occurring in normal peak traffic.

7.4 Other Fibre Channel diagnostic tools

All FC devices come with some kind of diagnostics, which should be used to troubleshoot problems and usually they are described in the devices' product manuals.

There is a saying: “physician, heal thyself” (it may not help too much being a doctor when you are the one who is sick!). Or, to put this another way, how much trust should you place in the failure indications reported by a failing device?

To find out what is really going on, your SAN service provider should be able to delve into the bits and bytes of the Fibre Channel data exchange when looking for protocol violations, time-out problems, and the kinds of failures that may corrupt your data — but that do not typically write to a log. The service provider will do this by using a Fibre Channel analyzer.

There are diverse Fibre Channel protocol analyzing tools (trace tools) on the market, which are able to collect in real-time Fibre Channel frames on 1 Gb/s and 2 Gb/s links, and optionally stop the trace on a triggered event and decode Fibre Channel and upper-layer protocols.

Notes® on preventive action: Look at assigning prepared points for Fibre Channel data interception in your cabling: patch panels and some of the switches provide with splitter or mirror options, so a protocol analyzer can be attached at any time without taking the fiber optic link offline.

The trace tools are designed for the needs of Fibre Channel protocol specialists to look into the details of data exchange and then to determine the cause of a malfunction.

Figure 7-7 and Figure 7-8 show typical decoding screens of two of the most popular Fibre Channel trace tools.

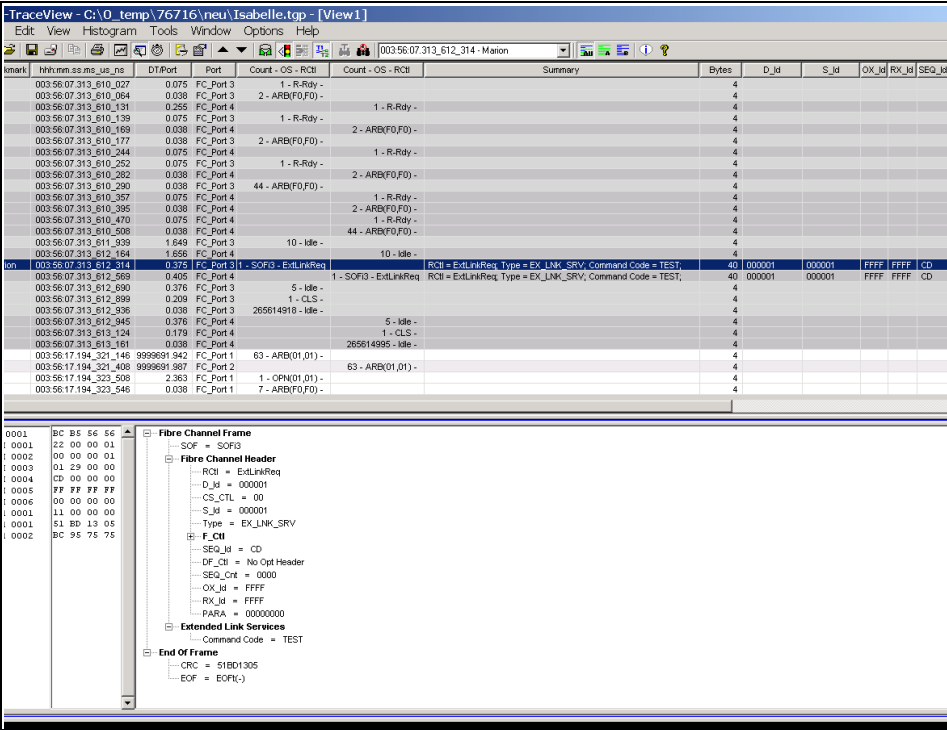


Figure 7-7 FINISAR GTX TraceView

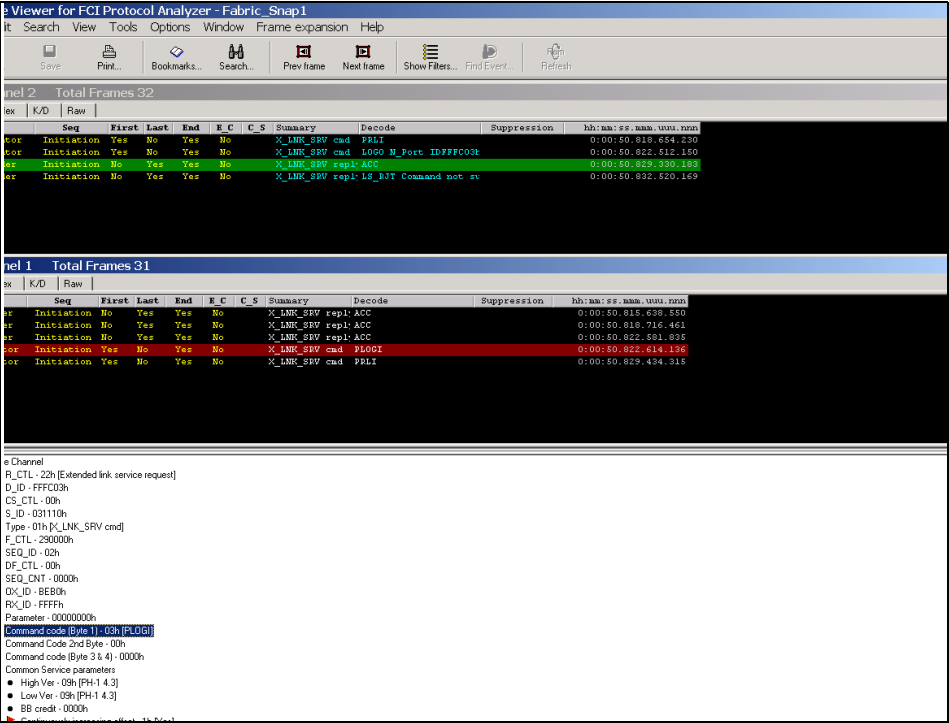


Figure 7-8 Xyratex FCI Protocol Analyzer

Although most of these trace tools offer expert functions, which go through the whole trace data and search for FC protocol violations and obvious matters between the FC devices like open requests, the tool itself is not always able to find the problem on its own. It still requires a specialist to analyze it.

More information about these vendors trace tools can be found at these Web sites:

<http://www.finisar.com/home/>
<http://www.xyratex.com/>



Part 2

Survival solutions

“My mind rebels at stagnation. Give me problems, give me work, give me the most abstruse cryptogram, or the most intricate analysis, and I am in my proper atmosphere. I can dispense then with artificial stimulants. But I abhor the dull routine of existence. I crave for mental exaltation.”

Sherlock Holmes.

Not wishing to let our minds stagnate, in the second part of this book, we offer some SAN survival solutions. Obviously it is not possible for us to describe every possible solution that exists; however, we will show how we have employed the components, features, and disciplines to suit our environment.



General solutions

In this chapter we discuss some of the building blocks and general concepts for building reliable and powerful SANs. Included are requirements for servers and storage and their software, as well as fabric devices.

We present various implementations and uses of a SAN environment, starting with a simple fabric and building it up into a complex design.

8.1 Objectives of SAN implementation

To ensure the highest level of system uptime, utilization, and security, companies are implementing reliable storage networks capable of boosting the availability of data for all the users and applications that need it. These companies typically represent the industries that demand the highest levels of system and data availability — the utilities and telecommunications sector, brokerages and financial service institutions, and a wide variety of service providers.

By reducing or eliminating single points of failure in the enterprise environment, SANs can help to improve overall availability of business applications. By utilizing highly available components and secure solutions as well as a fault-tolerant design, enterprises can achieve the availability needed to support 24x7 uptime requirements.

In vital networks such as SANs, with their associated hosts, fabric, and storage components, as well as software applications, downtime can occur even if parts of the system are highly available or fault tolerant. To improve business continuance under a variety of circumstances, SANs can incorporate redundant components, connections, software, and configurations to minimize or eliminate single points of failure.

Implementing multiple levels of redundancy throughout a SAN environment can reduce down-time by orders of magnitude. For instance, hardware components, servers, storage devices, network connections, and even the storage network itself can be completely redundant. A fundamental rule for improving fault tolerance is to ensure multiple paths through separate components regardless of a vendor's assurances of high availability. This is especially true when physical location and disaster tolerance are concerns, or when a complex device can become a single point of failure.

8.2 Servers and host bus adapters

To ensure availability, hosts should include redundant hardware components with dual power supplies, dual network connections, and mirrored system disks typically used in enterprise environments. Hosts should also have multiple connections — two independent connections is a minimum — to alternate storage devices through Fibre Channel switches. In most cases, servers should feature dual-active or hot-standby configurations with automatic failover capabilities.

8.2.1 Path and dual-redundant HBA

The next single point of failure to consider after the server is the path between the server and the storage. Potential points of failure on this path might include HBA failures, cable issues, fabric issues, or storage connection problems. The HBA is the Fibre Channel interconnect between the host and the SAN (replacing the traditional SCSI card for storage connectivity). Using a dual-redundant HBA configuration helps ensure that a path is always available. In addition to providing redundancy, this configuration will enable overall higher performance due to the additional SAN connectivity.

8.2.2 Multiple paths

To achieve fault tolerance, multiple paths are connected to alternate locations within the SAN or even to a completely redundant SAN. Server-based software for path failover enables the use of multiple HBAs, and typically allows a dual-active configuration that can divide workload between multiple HBAs — improving performance. The software monitors the “health” of available storage, servers, and physical paths and automatically reroutes data traffic to an alternate path if a failure occurs.

Path failover

In the event of an HBA or link failure, the host software detects that the data path is no longer available and transfers the failed HBAs workload to an active one. The remaining HBA then assumes the workload until the failed HBA is replaced or the link is repaired. After identifying failed paths or failed-over storage devices and resolving the problem, the software automatically initiates fail back and restores the dual path without impacting applications. If desired, an administrator can manually perform the fail back to verify the process.

The software that performs this failover is typically provided by system vendors, storage vendors, or value-added software developers. Software solutions, such as IBM Subsystem Device Driver (SDD), help ensure that data traffic can continue despite a path failure. These types of software products effectively remove connections, components, and devices as single points of failure in the SAN to improve availability of enterprise applications.

To help eliminate unnecessary failover, the software distinguishes between actual solid failures and temporary network outages that might appear to be solid failures. By recognizing false failures, the software can help prevent unnecessary failover/fallback effects caused by marginal or intermittent conditions. After detecting an actual failure, the software typically waits to determine whether the event is an actual failure.

The typical delay in the failover process can range from an instant failover (when a loss of signal light is detected) up to a minute (if the light signal is still available and the path failure is in another part of the network). These delays are typically adjustable to allow for a variety of configurations and to allow other, more rapid recovery mechanisms such as path rerouting in the SAN.

8.3 Software

One of the keys to improving availability is shifting the focus from server availability and recovery to application availability and recovery. Mission-critical applications should be supported on clustered or highly available servers and storage devices to ensure the applications' ability to access data when they need it — even in the midst of a failure. Sophisticated software applications can enable application or host failover, in which a secondary server assumes the workload if a failure occurs on the primary server. Other types of software, such as many database applications, enable workload sharing by multiple servers — adding to continuous data availability where any one of several servers can assume the tasks of a failed server.

In addition, many server vendors and value-added software providers offer clustering technology to keep server-based applications highly available, regardless of individual component failures. The clustering software is designed to transfer workload among active servers without disrupting data flow. As a result, clustering helps companies guard against equipment failures, keep critical systems online, and meet increased data access expectations.

Some clustering software, such as VERITAS Cluster Server, enables application failover on an application by application basis. This capability enables administrators to prioritize the order of application failover. Fibre Channel SANs facilitate high-availability clustering by simplifying storage and server connectivity. Moreover, SANs can provide one of the most reliable infrastructures for server clustering, particularly when clustered servers are distributed throughout the enterprise to achieve higher levels of disaster tolerance, a practice known as “stretched clusters.”

8.4 Storage

To improve performance and fault tolerance, many of today's storage devices feature multiple connections to the SAN. Multiple connections help guard against failures that might result from a damaged cable, failed controller, or failed SAN component, such as an SFP optical module. The failover process for storage connections typically follows one of the following methods.

Transparent failover

One method is transparent failover, in which a secondary standby connection comes online if the primary connection fails. Because the new connection has the same address as the original failed connection, failover is transparent to the server connection, and application performance is not affected. After the primary connection is repaired, it assumes the workload.

Active connections

Another method is to use dual or multiple active connections with each connection dedicated to certain logical volumes within a given storage system. If one connection fails, the other active connections automatically assume its logical volume workload until it comes back online. During this time, the alternate connections support all logical volumes, so there might be a slight performance impact depending on workload and traffic patterns.

Load balancing connections

A third method used for storage path failover also utilizes dual or multiple active connections. In this case, however, both connections can simultaneously access the logical volumes. This design can improve performance through load balancing, but typically requires host-based software.

During a storage connection failure, the alternate active connection continues to access the logical volumes. After the failed connection is repaired, the other path becomes active and load balancing resumes.

All of these failover methods are designed to ensure the availability of the enterprise applications that use them. In addition, failover generally is coordinated with server software to ensure an active path to data, transparent to the application.

Mirroring

Another effective way to achieve high availability in a SAN environment is by mirroring storage subsystems. SANs enable efficient mirroring of data on a peer-to-peer basis across the fabric.

These mirroring functions contribute tremendous fault tolerance and availability characteristics to SAN-based data. Combining the mirroring functions with switch-based routing algorithms (which enable traffic to be routed around path breaks within the SAN fabric) creates a resilient, self-healing environment to support the most demanding enterprise storage requirements. The mirrored subsystems can provide an alternate access point to data regardless of path conditions.

A common use of mirroring involves the deployment of remote sites within the enterprise. Implementing SANs through Fibre Channel switches enables the distribution of storage and servers throughout a campus, metropolitan area, and beyond. Fibre Channel overcomes many of the distance limitations of traditional SCSI connections, enabling devices to be extended over much longer distances for remote mirroring, tape backup, and disaster recovery operations.

8.5 Fabric

The switched fabric, as the central part of the SAN, is the focus of any discussion about performance and availability. The fabric design should provide a high performing environment for all storage-related enterprise applications and ensure connectivity even during partial outages. By implementing redundancy, the fabric design helps to prevent isolated failures from causing widespread outages and minimizes disruption to system operations.

8.5.1 The fabric-is-a-switch approach

Typically, when one thinks of a director, the assumption is that all the fabric redundancy is consolidated into one box. Theoretically, a director is supposed to provide full internal redundancy. Numbers of critical field replaceable units (FRUs) installed in a director will failover automatically should a component malfunction.

High availability is provided through the hardware and options, such as these:

- ▶ Redundancy of all active components
- ▶ All active components providing support for automatic failover
- ▶ Redundant power and cooling
- ▶ Hot swapping of all FRUs
- ▶ Hot swapping of spare ports
- ▶ Automatic fault detection and isolation
- ▶ Non-disruptive firmware updates

An example of this approach is shown in Figure 8-1.

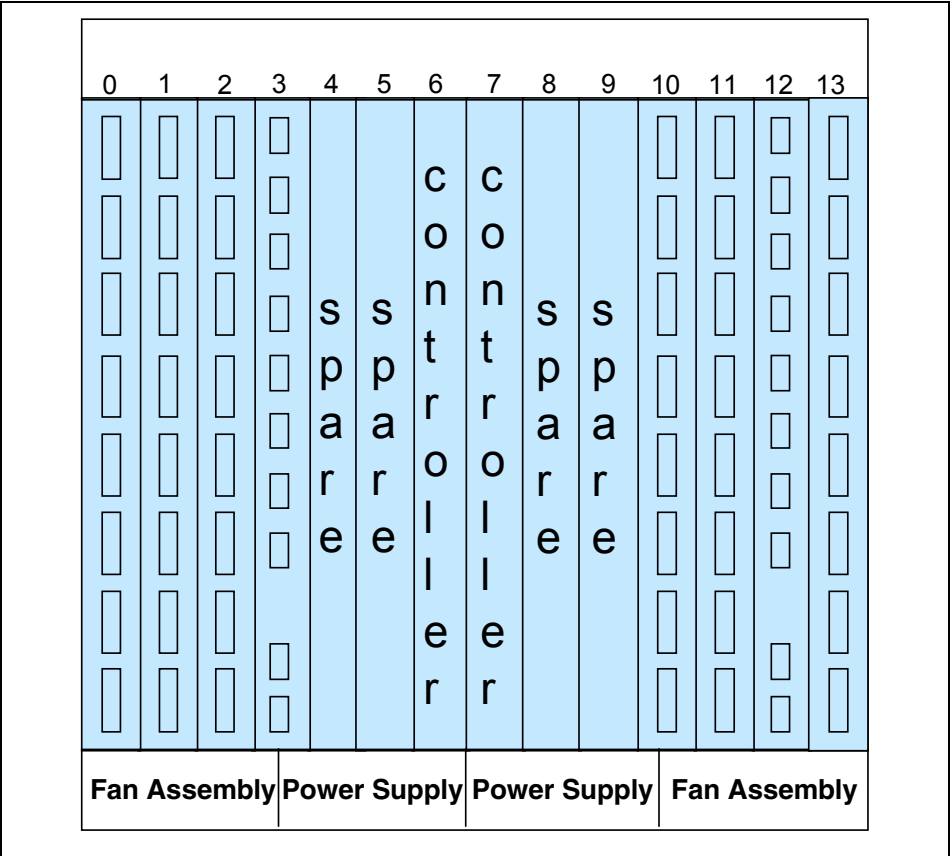


Figure 8-1 Fabric in a director

A director also has a high port density in a single footprint and can usually scale up to an even higher port count. You may build your fabric by implementing one director and have a highly performing and highly available fabric. From a security point of view, a single director is easier to handle and to protect than a widespread fabric, but there is still one single point of failure left, which is the fabric (director) itself. Intentionally or by user error, a fabric can be taken down and therefore the fabric or director should have a backup of its own: a dual fabric.

8.5.2 The fabric-is-a-network approach

Redundancy in the SAN fabric can be built through a network of switches to provide a robust mission-critical SAN solution. With its connected servers, switches, and storage ensuring high availability, the meshed fabric provides a most resilient infrastructure. With an infrastructure of switches, SAN administrators will scale their network to guarantee performance, availability and security by building it into the network rather than relying on a single footprint.

SAN infrastructures require high availability and a high port aggregation to solve problems such as backup and storage consolidation. Since ISLs can be utilized most efficiently nowadays, a network of smaller switches may enable the SAN to support the appropriate level of bandwidth by increasing the number of switches and can be considered as an opposite strategy to the pure director-based SANs.

However, there is no design that is without a limit. To provide every server in the SAN with the appropriate bandwidth to exchange data with its storage, all at the same time, is not a feasible concept. It would mean having to provide ISL bandwidth for the cumulative server port bandwidth, or in other words to give each server its own dedicated ISL — but that is not in the spirit of networking. Ideally, a network should be oversubscribed to the maximum point possible, while maintaining the minimum acceptable performance.

This approach ensures that the fewest resources can support the greatest numbers of users or applications. A typical oversubscription ratio will be 7:1 or higher to start with. During operation you will observe port performance and decide whether to implement more ISLs or more device ports. By taking advantage of the scalability of the SAN switches and ISL trunking features, the switched fabric can be tailored very easily at the first and subsequent implementations.

8.6 High level fabric design

You can configure scalable solutions that help address your needs for high performance and availability for environments ranging from small workgroups to very large, integrated enterprise SANs.

If you start with a single switch, you will find that when your fabric grows, you will need to connect new switches to it. The first step may be a cascaded design. We show two possible options in Figure 8-2.

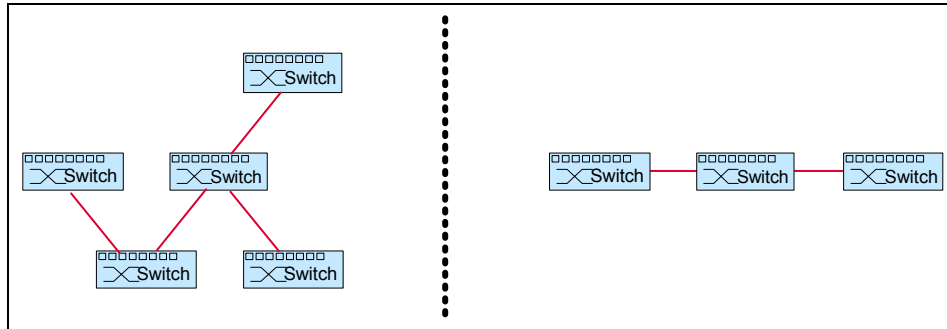


Figure 8-2 Two examples of switch cascading

When cascading switches you will need $(n-1)$ -ISLs to connect n -switches. It raises your port rate compared to using one switch. With a cascaded fabric it is possible to introduce a bottleneck when traffic has to travel down the ISLs. For this reason there are many ways to ensure that you do not introduce a bottleneck into the SAN.

A next step towards higher performance and higher availability is a ring design, as shown in Figure 8-3.

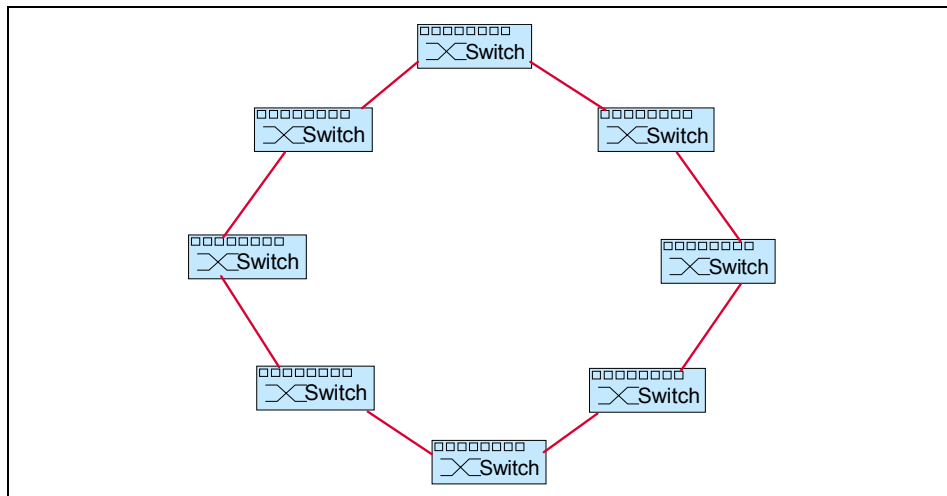


Figure 8-3 Ring design

Whenever one ISL fails, there is still connectivity throughout the whole fabric. However, if an ISL fails, then it may take more hops for the initiator to reach the target. To connect n -switches, you'll need n -ISLs.

However, these SAN designs do not show very much thought or structure to support the traffic flow. We could dedicate switches to be connected to storage or host only, but all the traffic would have to pass through the ISLs and this may be counter-productive. Also, simply structuring the SAN by dedicating switches to diverse departments will not increase performance nor availability.

The way to increase performance and availability in a fabric is to build a network of switches in a meshed network topology, as shown in Figure 8-4.

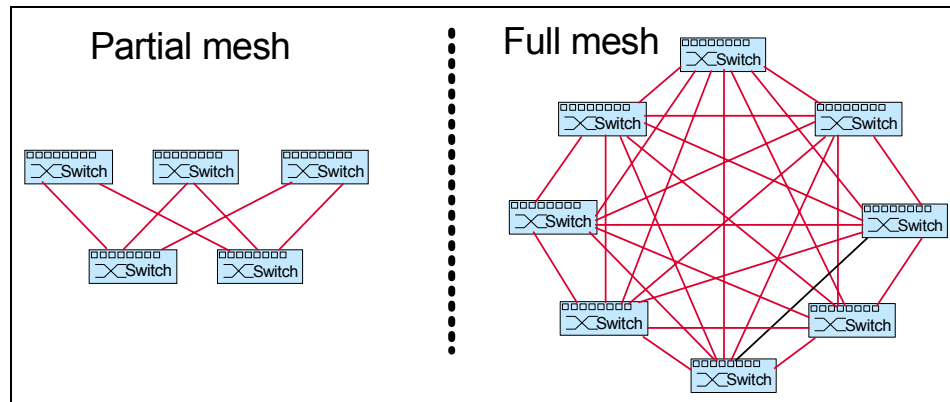


Figure 8-4 Meshed network design

Figure 8-4 shows a partial mesh and full mesh fabric design. For a partial mesh fabric, you will need at least n -ISLs to connect n -switches. For a full mesh fabric, you will need m -ISLs to connect n -switches, as follows:

$$m = \frac{n^2 - n}{2}$$

The number of ISLs increases very quickly when the count of switches increases. A full mesh network offers the highest performance and availability.

With either type of meshed topology, it is easy to structure the fabric for the sake of easier maintenance and administration, fault isolation, and higher traffic flow.

A common structure is a tier-layer design with a dedicated layer of switches for hosts and a layer for storage devices, as shown in Figure 8-5.

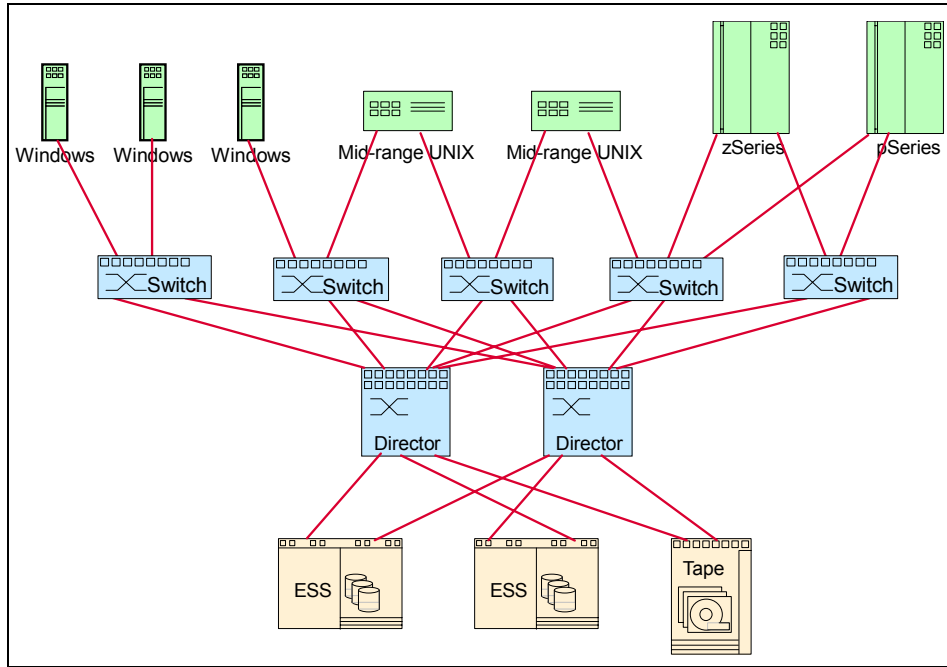


Figure 8-5 Host-tier and storage-tier

It is a partial mesh design with all hosts connected to the upper tier, and all storage devices to the lower one. Every data transfer from host to storage will cross the ISLs and we have to keep that in mind when provisioning the ISLs.

If the SAN were to grow bigger and there were eight switches in each tier, in order to connect every switch in the host-tier to every switch in the storage tier, this would cost 128 ISL ports, as shown in Figure 8-6.

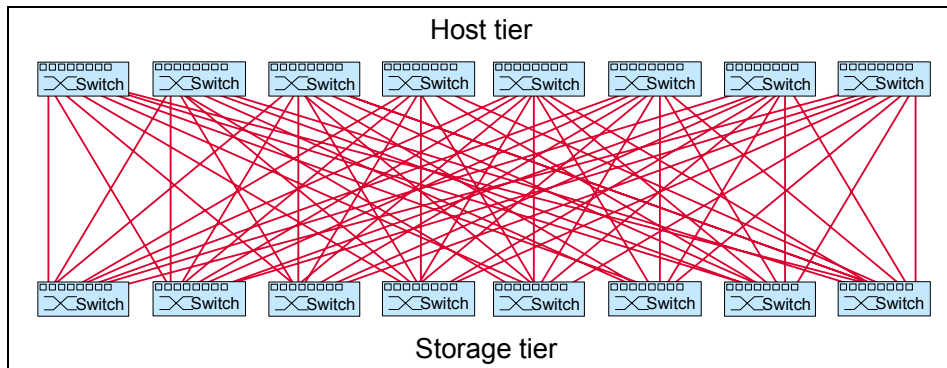


Figure 8-6 Tier to tier

This is just done to single-connect the switches from one tier to the other. The fabric does not gain any higher availability by such a design.

A better way to connect the tiers would be to introduce a “focal point” for all ISLs between the tiers, called core switches; the switches at the host and storage tier are called edge switches. We now have a core-edge design like that shown in Figure 8-7.

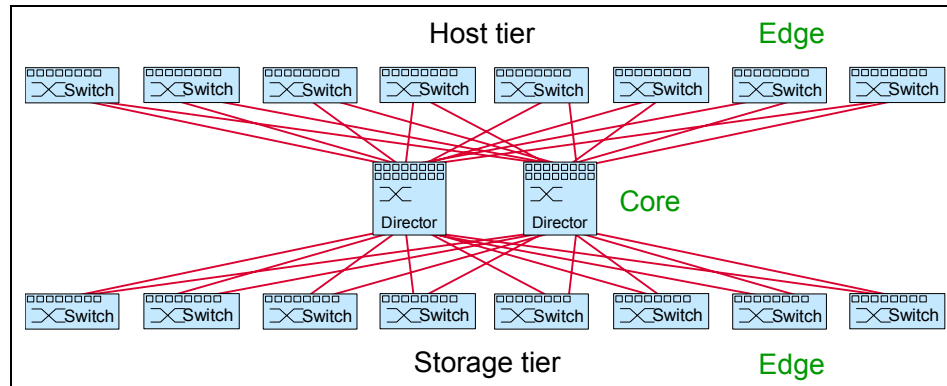


Figure 8-7 Core-edge design

With the core-edge design, you will have any-to-any edge-switch connectivity without having to connect any-to-any switch. We do have less cumulative ISL bandwidth in the SAN now, when compared to the design in Figure .

As the core is the focal point, you will want to deploy your core switches or directors that have redundancy inherent in their design. A storage-tier usually needs less ports than a host-tier. So, in some cases, when your storage devices are pooled locally, you don't build up a separate edge-tier for storage, as shown in Figure 8-7. Rather, the ports of the core switches connect to the storage devices directly, as shown in Figure 8-5.

8.7 Definitions

In our examples in this chapter and in the following chapters, we will use the following terms:

- **Oversubscription:** This term means the ratio of the number of input devices weighted by their individual bandwidth to the number of output devices also weighted by their individual bandwidth. That may be the amount of hosts connecting to a storage device. For example, a storage device can handle up to 100 MB/s on one port, connecting four servers which will do 60 MB/s each will give us an oversubscription of 2.4:1:

$$oversubscription = \frac{\sum port_{input} \times bandwidth}{\sum port_{output} \times bandwidth} = \frac{4 \times 60 \text{ MB/sec}}{100 \text{ MB/sec}} = \frac{24}{10}$$

- **ISL-oversubscription:** This special case of oversubscription takes the ratio of host ports to the possible ISLs carrying that traffic - again we take bandwidth of the individual ports into account. ISL-oversubscription is of interest in a meshed fabric. The higher the ratio is, the more devices will share the same ISL and the more is it likely we will suffer from congestion. Adding a 2 Gb/s ISL (that is 200 MB/s) to our previous example will give us a value of 1.2:1:

$$\text{ISL-oversubscription} = \frac{\sum port_{input} \times bandwidth}{\sum ISL \times bandwidth} = \frac{4 \times 60 \text{ MB/sec}}{200 \text{ MB/sec}} = \frac{240}{200}$$

Note: If all ports on the switches are operating with the same speed, it is a simple division to calculate ISL-oversubscription. In cases where ports with 2 Gb/s and 1 Gb/s are intermixed, each host port and each ISL has to be multiplied by its bandwidth before computing the ratio.

- **Fan-out:** This is the ratio of server ports to a connected storage port. Fan-out differs from oversubscription in that it represents a ratio based on the number of connections regardless of the throughput. Our previous example would result in a fan-out of 4:1:

$$\text{Fan-out} = \frac{\sum port_{host}}{\sum port_{storage}} = \frac{4}{1}$$

- **Topology:** This is a synonym for design in networking. Fibre Channel historically supports only three topologies: point-to-point, arbitrated loop, and switched fabric, but this is *not* what is meant here. Topology in our solutions means the design itself of the switched fabric and whether it is a cascaded, meshed, tiered, or a core-edge design.

8.7.1 Port formulas

As we have stated already, oversubscription is an accepted bottleneck. Depending on the load profile of your Fibre Channel devices, you will accept these bottlenecks in such a way that they will never allow the SAN to end up in gridlock.

Typically the first implementation of a SAN is based on assumptions. By using this port formula, you may estimate how many host ports you will get out of a two-tier fabric with a given number of switch ports, and estimated values for host-to-storage oversubscription ($over_{hs}$) and ISL-oversubscription ($over_{ISL}$):

$$\sum port_{host} = \frac{\sum port_{fabric} - \sum port_{spare}}{1 + \frac{1}{over_{hs}} + \frac{2}{over_{ISL}}}$$

As an example, we will use the diagram Figure 8-5 on page 279. We will take seven 32-port switches and assume that we want to use only 80% of the ports in our first implementation, saving the other 20% for future expansions. We will assume a host-to-storage oversubscription of 6:1 and an ISL-oversubscription of 10:1. This is shown in Example 8-1.

Example 8-1 Host-to-storage oversubscription

5 switches with 32 ports each: $port_{fabric} = 160$
20% spare ports: $port_{spare} = 32$
oversubscription host-to-storage: $over_{hs} = 6:1$
ISL-oversubscription: $over_{ISL} = 10:1$

$$\sum port_{host} = \frac{\sum port_{fabric} - \sum port_{spare}}{1 + \frac{1}{over_{hs}} + \frac{2}{over_{ISL}}} = \frac{160 - 32}{1 + \frac{1}{6} + \frac{2}{10}} = 93$$

So:

The host ports come to: $port_{host} = 93$

The storage ports will be $port_{storage} (93 / 6) = 15$

The ISLs will be $(93 / 10) = 10$, which is 20 ISL ports

The assumptions about oversubscription in the previous example will delegate the used ports of the fabric to 73% as host ports, 12% as storage ports, and 15% as ISL ports.

You can use that formula for a core-edge design too (see Figure 8-7 on page 280). If you ignore the core ports and just count the ports of your edge-switches to get $\Sigma \text{port}_{\text{fabric}}$, simply calculate the other port counts and assume that your core ports are the same value as the ISL ports. So, referring back to Example 8-1 on page 282, your core would consist of another 20 ISL ports.

8.8 Our solutions

In the following chapters we categorize and discuss the relevant items of SAN design in various implementations with different switches and directors. We have categorized the solutions according to:

- ▶ Performance
- ▶ Availability
- ▶ Distance
- ▶ Clustering
- ▶ Secure

We have also added any vendor specific and unique solutions. Although we categorize, for example, a solution as a *performance* solution, this does not mean that it is only a performance solution. It will typically contain elements of all the other categories as well, but we have just chosen to focus on one aspect for clarity. A *Checklist* and “*What If*” *failure scenario* complements each solution.



Cisco director and switch solutions

In this chapter we illustrate and describe solutions based on the Cisco MDS 9000 family of switches and directors.

The Cisco MDS 9509 Multilayer Director is the IBM 2062-D07 (which has seven optional slots for switching modules).

The MDS 9216 Multilayer Switch is the IBM 2062-D01 (which has one optional slot for a switching module). The solutions are categorized as follows:

- ▶ Performance solutions
- ▶ Availability solutions
- ▶ Distance solutions
- ▶ Clustering solutions
- ▶ Secure solutions
- ▶ Loop solutions

9.1 Performance solutions

When there is little or no server performance information available, it is very difficult to work out the ratio of server ports to storage ports. The solution we show in Figure 9-1 illustrates how a general high performance profile could be applied to a SAN design using a director and a single ESS.

This methodology should only be used to generate a high level design. Final designs must be based on performance data collected from the servers.

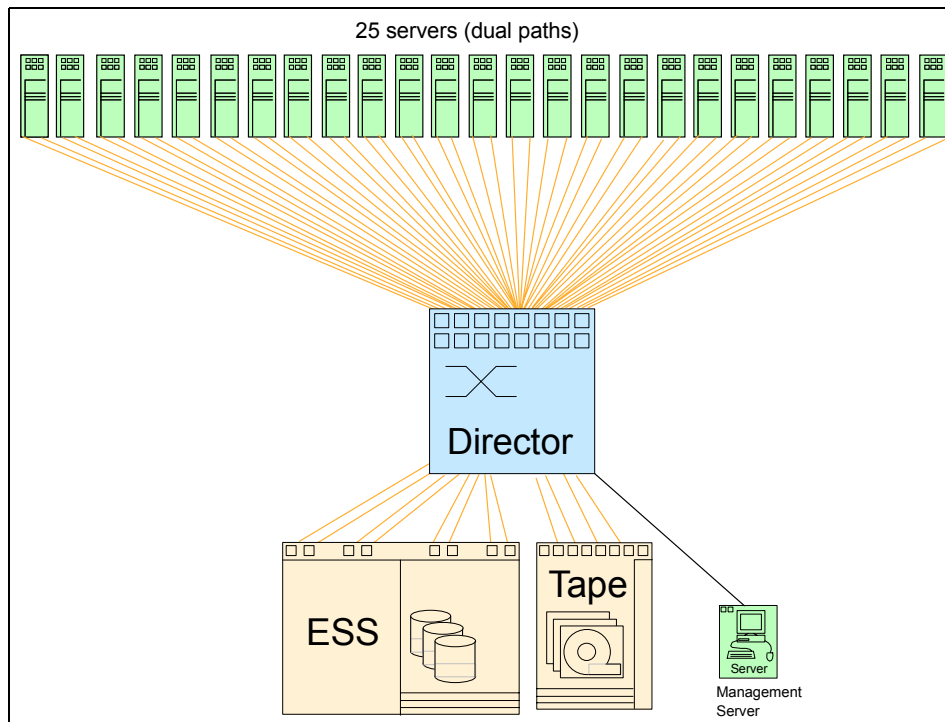


Figure 9-1 High performance design

Typically a fiber optic cable link to the ESS will operate at up to 130 MB/s. If we configured eight connections from the director to ESS, we would have a maximum SAN peak bandwidth capability of 1040 MB/s (8 x 130 MB/s).

If we connected 25 dual attached servers to the director and all servers were processing at the same time, we would potentially have a maximum SAN peak bandwidth to the ESS of 41.6 MB/s per server (1040 MB/s / 25).

This throughput assumes that all 25 servers are able to generate this level of I/O at the same time. This could be categorized as a high performance profile.

Based on this theory, for a *high performance profile*, we have a server connection to ESS port ratio of 6.25 which we round down to 6. So our ratio in this case is 6:1.

Note: The high performance profile is calculated by determining the ratio between the number of server ports (or HBAs) and ESS Fibre Channel ports.

In our example:

25 servers with dual paths = 50 server ports / 8 ESS ports = ratio of 6.25:1

For *low performance profiles*, such as file and print servers, we use a rule-of-thumb of 12 server connections to one ESS port. In this case we would use a ratio of 12:1.

Tape device functions such as server-less backup or LAN-free backup should also be taken into consideration for a complete SAN profile.

These profile ratios are recommended as a starting point when there are no server performance details available. These rules are very generic and should only be applied at the initial design stage. Prior to any final design a detailed performance profile should be conducted using open systems performance measuring tools such as IOMETER and IBM Disk Magic.

In our solution we will connect 25 dual attach high performance profile servers to a single ESS.

Components

- ▶ SAN fabric:
 - Cisco MDS 9509 Director configured with 64 ports (4 x 16 port switching modules)
- ▶ Servers:
 - 25 servers each configured with dual FC HBAs.
- ▶ Storage:
 - ESS-800 configured with 8 x FC adapters
- ▶ Tape:
 - 3584 Automated Tape Library configured with 6 x FC drives
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS
 - 3584 with StorWatch Specialist
 - IOMETER and Disk Magic for performance modelling.

Checklist

We checked the following items:

- ▶ Spread dual connected ports across blades to minimize the effect of a blade failure within the director.
- ▶ Spread ESS connections across switching modules in order to protect against a module failure.
- ▶ Consider the impact of losing a switching module and balance the server groups to minimize impact.
- ▶ Leave some ports spare for contingency.
- ▶ Monitor the performance of the environment with Cisco Fabric Manager.
- ▶ Collect MIB information to determine busy ports.
- ▶ Conduct a detailed server performance profile.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

As detailed in our solution description, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth. Due to the performance of the director, any SAN performance bottlenecks will likely be at the ISLs (if configured), or more likely at the HBAs of the storage device.

Based on this theory, the performance of the SAN will be determined on how much traffic will be moved through the E_Port or HBA. With detailed server profiles it is possible to balance this accordingly.

The Cisco MDS 9000 family of switches and directors provide unique flexibility in designing a performance based solution due to the underlying architecture of the equipment. For example, we could configure a director using 4 x 16-port switching modules to provide 64-ports with a full non-blocking implementation. Alternatively, if we know some servers do not require such high performance, as they are not able to generate such high bandwidth requirements, we could consolidate these servers onto the 32-port switching modules with shared bandwidth capability. This allows us to design SAN solutions on a performance basis while ensuring we get the best cost-per-port and best port density.

Scalability

Based on our performance profiling, we could expand our solution and connect two directors together using dual E_Ports, as shown in Figure 9-2. Each director now has 4 connections to the ESS and 3 connections to the tape library. We have now created a higher availability SAN that could support 100 device connections

(assuming 50 servers with dual HBAs) attached. This design provides protection against any possible failure of a director, such as a back-plane failure.

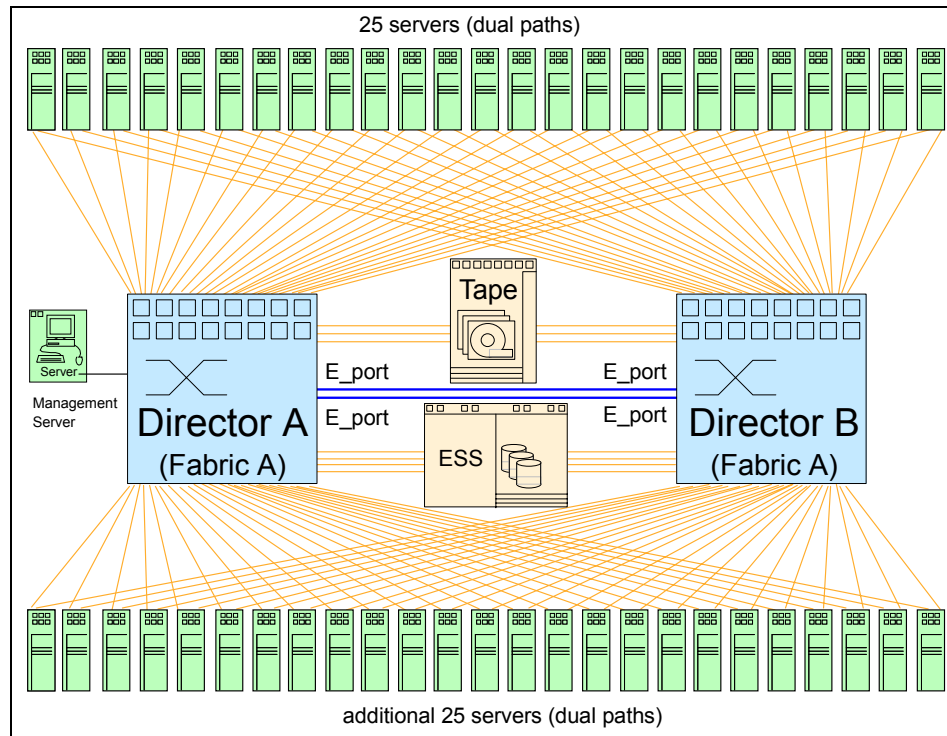


Figure 9-2 Expanding the SAN fabric via E_Ports

Servers have increased the server port to storage port ratio to 12:1, but reduced the maximum SAN server bandwidth to 10.4 MB/s per server. This design is a much more cost effective solution. The solution could be scaled up through the addition of extra switches or directors.

Availability

While this design provides a higher-availability design than for the single director model a failure in the SAN fabric could result in all hosts losing access to the devices. For example, if an invalid zoning change was made to the fabric or the fabric was corrupted this would affect all devices in the SAN.

The Cisco MDS 9000 family provides Virtual Storage Area Networks (VSAN) technology which is a way of creating totally independent fabrics across the same hardware platform. We will illustrate solutions based on VSAN technology in the solution in 9.2, "Availability solutions" on page 290.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

These are some theoretical assumptions based on Figure 9-1:

- ▶ Spread the servers with the high performance profile across 16-port switching modules. A single 16-port switching module will provide 2 Gb/s throughput to all ports on the module whereas the 32-port switching module provides shared bandwidth to the port group of 4 Gb/s.
- ▶ If a cable fails between the director and ESS, an alternate route will be used, although we would lose 12.5% of the total available bandwidth.
- ▶ Regarding the director, if a switching module fails we still have connectivity, as we have dual connections, but we would lose 50% bandwidth to any connected servers, and up to 12.5% bandwidth to the ESS.
- ▶ If a supervisor module fails, there would be no effect as the spare supervisor module will automatically take over.
- ▶ If the backplane was damaged, we would lose connectivity to all servers at that site. The solution in Figure 9-2 provides protection against a backplane failure.
- ▶ If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- ▶ If an ESS HBA is unavailable we have multiple other connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- ▶ If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay. The ESS has 4 adapter bays each holding up to 4 interface cards. In our example we have configured the ESS with 2 x FC adapter cards per bay.
- ▶ For both Figure 9-1 and Figure 9-2, a failure in the SAN fabric itself will cause a loss of connectivity for all devices. We can use Cisco's Virtual SAN Technology to protect against such a failure and this is discussed in the topics that follow.

9.2 Availability solutions

Continuing on from the solution design in the previous section, we now look at two solutions aimed at providing the highest possible availability. These may not

be applicable to all environments, but they illustrate the issues associated with designing a highly available SAN infrastructure.

9.2.1 Dual fabric

One of the issues we have mentioned previously is that a failure in the SAN fabric can cause the entire SAN to become unstable. A single SAN fabric could be affected by a number of events including, but not limited to, the following:

- ▶ Incorrect zoning change
- ▶ Overlaying of a zone configuration
- ▶ Broadcast storm
- ▶ SAN segmentation (occurring when a switch is incorrectly attached to an existing switch with the same domain ID)

By implementing a solution based on dual fabrics, we can avoid the impact of a SAN fabric failure. Such a solution is described in Figure 9-3. In this scenario, every device in the SAN has a connection to the separate fabrics, and this is how a traditional SAN would be designed to protect against a single fabric failure.

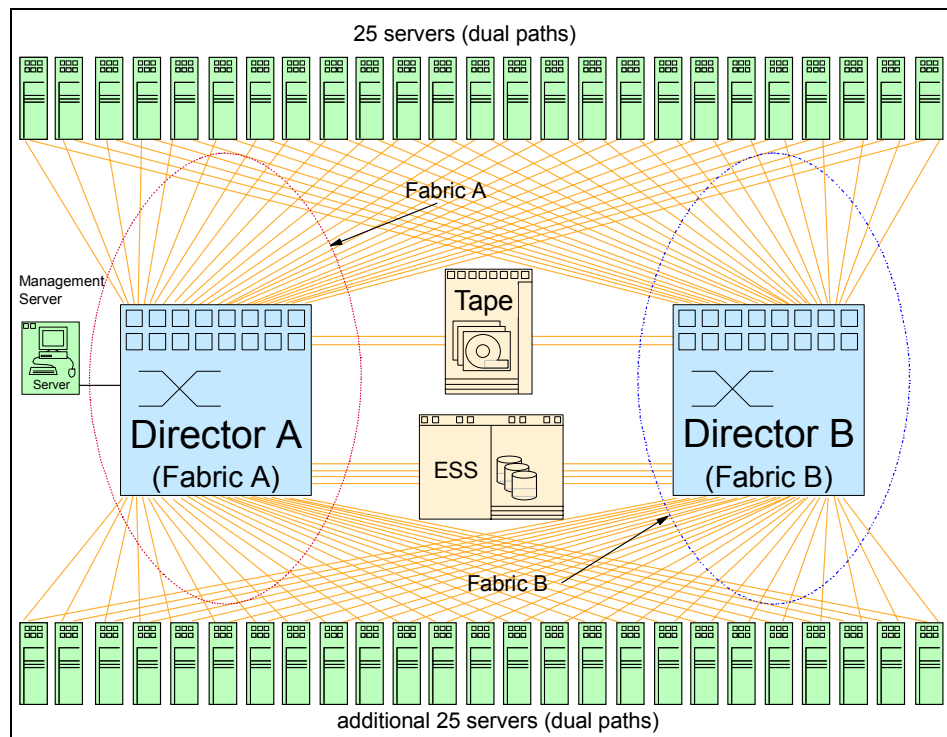


Figure 9-3 Traditional dual fabric design (without VSANs)

By using Cisco's Virtual SAN (VSAN) technology, we eliminate the requirement for the second director, as shown in the previous example. Shown in Figure 9-4 is an illustration of how we can provide dual SAN fabrics providing the same logical infrastructure as in the previous diagram. We have illustrated this with 25 servers, but this could be many more.

The diagram shows how two distinct VSANs are formed using the one physical infrastructure. The VSANs are logically separate from each other. Note that with VSANs it is not possible to have the same port defined in more than one VSAN.

This technology enables us to provide a dual-fabric solution without having to purchase multiple directors or switches to achieve this.

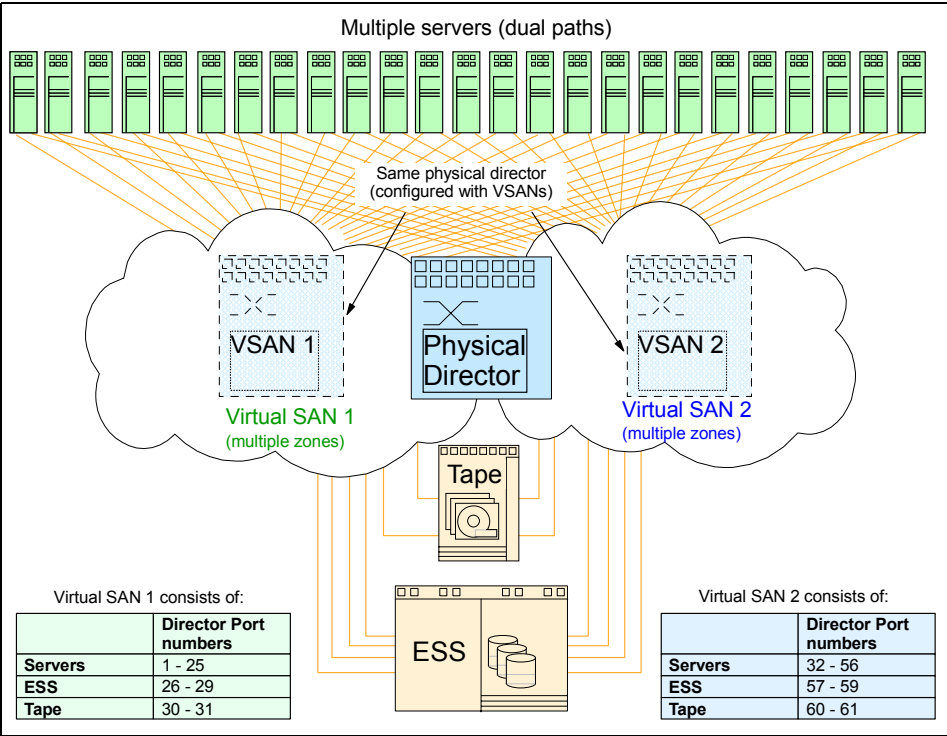


Figure 9-4 Dual fabric design with VSANs

Components

- ▶ SAN fabric:
 - 1 x Cisco MDS 9509 Director configured with 64 ports each (using 4 x 16-port switching modules).
- ▶ Servers:
 - Multiple servers each configured with dual FC HBAs.
- ▶ Storage:
 - 1 x ESS-800 configured with 8 x FC adapters.
- ▶ Tape:
 - 1 x 3584 Automated Tape Library configured with 4 x FC drives.
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS
 - 3584 with StorWatch Specialist
 - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- ▶ Set up VSANs ensuring that no port is in more than one VSAN.
- ▶ Ensure that all devices with multiple paths (servers, ESS, tape) have a connection to both VSANs.
- ▶ Spread dual connected ports across the switching modules.
- ▶ Spread ESS connections across switching modules to minimize any impact should a module fail.
- ▶ Consider the impact of losing a switching module and balance the server groups to minimize impact.
- ▶ Leave some ports spare for contingency.
- ▶ Collect MIB information to determine busy ports.
- ▶ Conduct a detailed server performance profile.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

As detailed in our solution description in 9.1, “Performance solutions” on page 286, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN’s bandwidth.

It is important to monitor the links into the tape library and the ESS as more hosts are connected to the SAN, in order to ensure that the links into the devices do not become saturated as more and more load is placed on the SAN.

Scalability

Based on our performance profiling, we could expand our solution further through the addition of extra directors into each VSAN. In this case we would connect the directors within the fabric together using TE_Ports, but we would not join the fabrics. TE_Ports enable us to aggregate the bandwidth across the ISLs.

Keeping some spare ports available in the event of a failure is a good practice if it is practical in your environment. This allows for the re-cabling or re-connecting of devices immediately in the event of a SFP or switching module failure.

Availability

While this design provides a higher-availability design than for the single SAN fabric model, it does not protect us against a site failure such as those discussed in 1.2, “Using a SAN for business continuance” on page 8.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

These are some theoretical assumptions based on Figure 9-3.

- ▶ If all high performance profile servers are on the same switching module, the director is a non-blocking device, so there should be no performance impact, although it would be sensible to spread the load.

If we were using 32-port switching modules, this would become a key design consideration as this module provides a total bandwidth of 32 Gb/s for the whole module or 200 MB/s of sustained bandwidth per port-group (each port group has 4-ports).

- ▶ If a cable fails between director and ESS, an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- ▶ Regarding the director, if a switching module fails, we still have connectivity as we have dual connections, but we would lose 50% bandwidth to any connected servers, and 12.5% bandwidth to the ESS.
- ▶ If a supervisor module fails, there would be no effect, as the spare supervisor module will automatically take over.
- ▶ If the director was physically damaged or destroyed, then obviously we would lose access to all the devices. This is one advantage of having redundant hardware in place.

- ▶ If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- ▶ If an ESS HBA is unavailable, we have multiple connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- ▶ If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay.
- ▶ Although both VSANs can be managed from one workstation, it will increase the number of zones and definitions that need to be managed by the administrator.

9.2.2 Dual fabrics over distance

To implement a dual fabric solution over distance would require multiple directors at the Primary and Secondary Sites, as shown in Figure 9-5. This design provides the highest availability, as we have dual fabrics separated over distance to a secondary site.

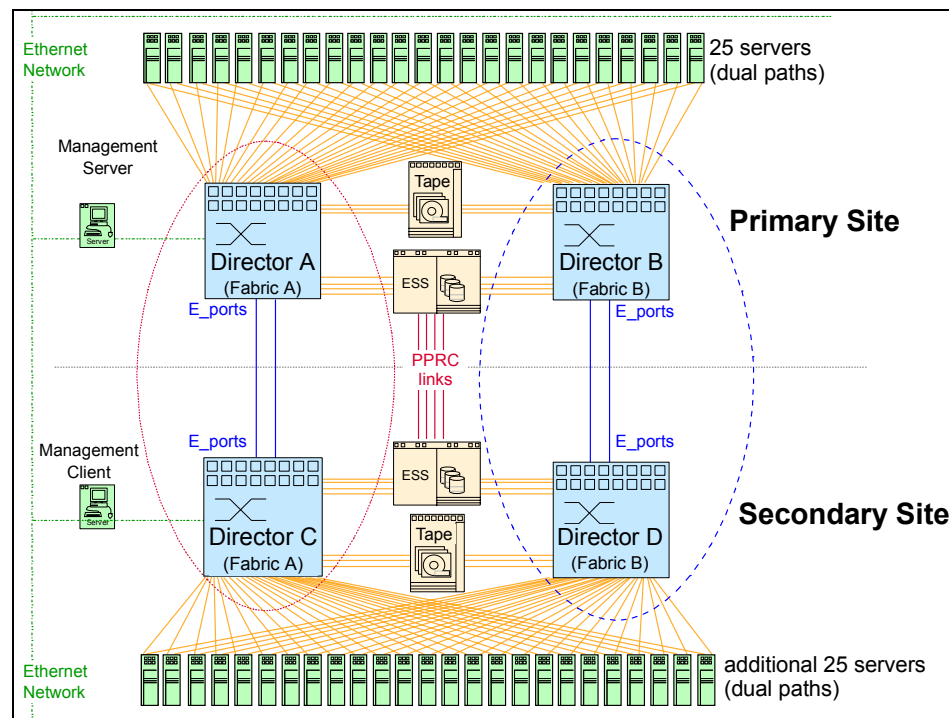


Figure 9-5 Traditional dual fabric design over distance

Using the Virtual SAN feature of the Cisco MDS 9000 family, we can put together a dual fabric and dual site design using half the number of directors when compared to the solution in Figure 9-5 on page 295.

The solution described in Figure 9-6 is designed to provide protection against a site failure.

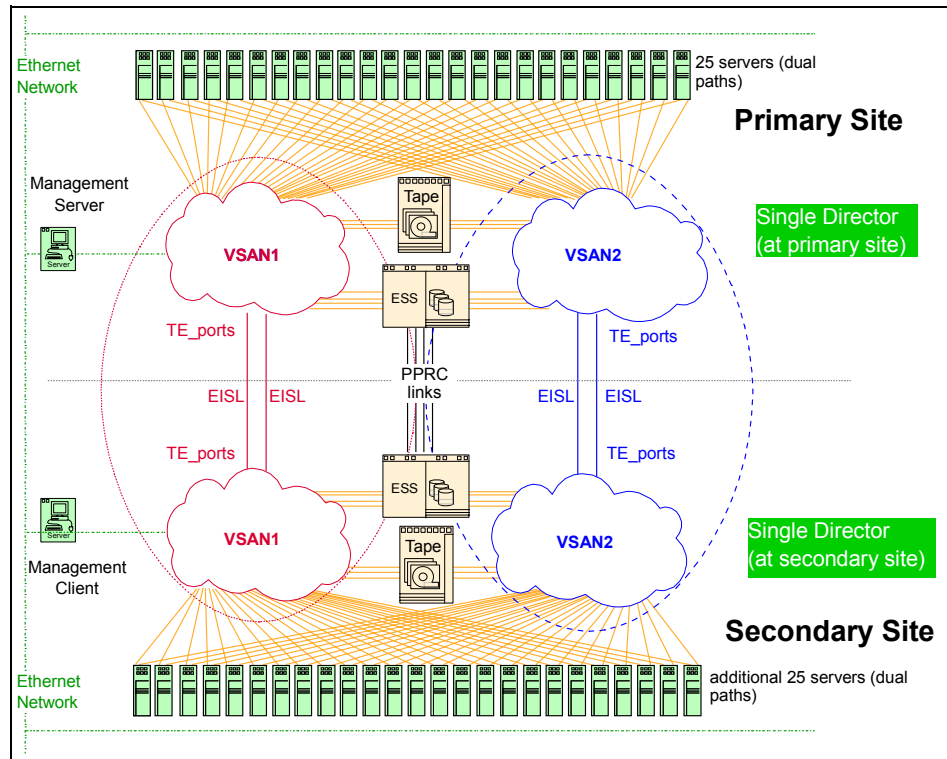


Figure 9-6 Dual fabric design over distance using VSANs

In this scenario, every device in the Primary Site has a connection to both VSAN1 and VSAN2 fabrics. The directors are connected via an EISL or TE_Port to a director located at the Secondary Site. The ESS's are connected using PPRC so all updates are mirrored between sites.

With this scenario, a failure at the Primary Site can cause our clustered servers to fail over to the Secondary Site. A VSAN fabric failure will cause a server to fail over to the surviving VSAN fabric. While performance could be affected, the servers would still have access to their data.

With this solution, the servers could be individual servers with a warm standby server at the remote location made possible via a manual failover process, or they could be clustered systems with an automated failover to a hot machine.

Components

- ▶ SAN fabric:
 - 2 x Cisco MDS 9509 Directors configured with 64 ports (including two longwave SFPs in each director for EISLs).
- ▶ Servers:
 - Multiple servers each configured with dual FC HBAs
- ▶ Storage:
 - 2 x ESS-800 configured with 8 x FC adapters for each site
 - PPRC license for remote copy between ESS's
 - ESCON adapter for Remote Copy
- ▶ Tape:
 - 2 x 3584 Automated Tape Library configured with 6 x FC drives for each site
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS
 - 3584 with Storwatch Specialist
 - IOMETER and Disk Magic for performance modelling.
- ▶ Cross site links:
 - It is assumed that the private FC links are available to go the required distance

Checklist

We checked the following items:

- ▶ Spread dual connected ports across switching modules.
- ▶ Spread ESS connections across switching modules to minimize any impact should a module fail.
- ▶ Consider the impact of losing a switching module and balance the server groups accordingly to minimize impact.
- ▶ Leave some ports spare for contingency.
- ▶ Collect MIB information to determine busy ports.
- ▶ Conduct a detailed server performance profile.
- ▶ Ensure that diverse routes are used for the PPRC and EISL (TE_Ports).

- ▶ Ensure that EISLs (TE_Ports) for each VSAN are located on different switching modules. This ensures that if a switching module should fail, there would still be an active path between directors.
- ▶ Ensure that EISLs and other devices requiring high performance are only attached using 16-port switching modules in order to maximize the available bandwidth.
- ▶ Ensure that any servers fitting the low performance profile are connected to the 32-port switching modules.
- ▶ Ensure that EISLs and PPRC links take diverse routes to ensure that they are not both affected by a cable failure.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

As stated in our solution description, a detailed server performance profile needs to be undertaken to be confident that there is no under- or over-utilization of the SAN's bandwidth.

It is important to monitor the links into the tape library and the ESS as more hosts are connected to the SAN. This ensures that the links into the devices do not become saturated as more and more load is placed on the SAN.

The performance of the ISLs will also need to be monitored over time to ensure that they are not overloaded. As the SAN increases it is possible that the bandwidth between the directors is no longer sufficient to meet the performance requirements. In this case, we would implement a PortChanneling solution to form an aggregated logic path across the ISLs. This can also be extended by adding another ISL to the directors to increase the total bandwidth available. This would be more of an issue if, for example, we were implementing a remote tape vaulting solution, in which case a significant amount of data would flow across the ISLs. Our advice is to monitor the ISL traffic over time and have an action plan ready to implement if the ISL links start to become saturated.

Scalability

Based on our performance profiling we could expand our solution further through the addition of extra directors into each site and expanding the ports in each VSAN. In this case we would connect the directors at each site using TE_Ports but we would continue to maintain the dual fabric through the VSAN technology.

Keeping spare ports available in the event of a failure is a good practice if it is practical in your environment. This allows for the re-cabling or re-connecting of devices immediately in the event of a switching module or SFP failure.

Availability

This design provides the highest possible availability but at the cost of having redundant directors and servers at the secondary site.

Security

We have not considered any security issues with this solution. These will be addressed in some of the following solutions.

“What If” failure scenarios

These are some theoretical assumptions based on Figure 9-1:

- ▶ If all high performance profile servers are on the same switching module, the director is a non-blocking device so there should be no performance impact, although it would be sensible to spread the load for availability reasons.
- ▶ If a cable fails between director and ESS an alternate route will be used, although we would lose 12.5% of the available bandwidth.
- ▶ Regarding the director, if a switching module fails, we still have connectivity as we have dual connections, but we would lose 50% bandwidth to any connected servers, and 12.5% bandwidth to the ESS.
- ▶ If a supervisor module fails, there would be no effect as the redundant supervisor module will automatically take over.
- ▶ In the event of a catastrophic failure of the director, the servers would fail over to the secondary site. This could be an automatic failure for clustered servers or a manual failover depending on the server requirements. Performance would not be an issue as the Secondary Site is configured to match the Primary Site requirements. This is not always the case with secondary sites so it is suggested that a performance analysis of the primary site is undertaken in order to ensure the secondary site can meet these requirements.
- ▶ If a server HBA fails, we lose up to 50% of the server's SAN bandwidth, and depending on the application, up to 30-40% of the server's performance.
- ▶ If an ESS HBA is unavailable, we have multiple connections that will automatically be used. We would lose 12.5% of the available SAN bandwidth per HBA.
- ▶ If an ESS bay is unavailable, we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth. The ESS is configured with 8 x FC adapter (2 in each bay) therefore if we lose one bay we would lose 2 x FC adapters.
- ▶ It is also important to consider the implications of fail-back procedures. Assuming the primary site failed over to the secondary site, all updates are now occurring at the secondary site. At some stage the data on the ESS at

the secondary site will need to be failed back to the ESS at the primary site to bring the data back in-sync. This may place increased load on the SAN infrastructure.

- ▶ In the event that a server needs to access the storage at the remote site this can put more pressure on the EISL bandwidth so this needs to be considered as part of the SAN planning process.
- ▶ More EISLs could be required if remote tape vaulting was to be implemented so this also needs to be factored in to the SAN design process.
- ▶ If remote tape vaulting was required it would be a useful idea to put these ports and a number of dedicated ISL's into a new VSAN only for tape activity. This would enable the organization to restrict or control the bandwidth that the tape drives were able to use and would prevent the tape activity from impacting on the production systems. By using VSANs we can ensure that traffic from one application does not affect another application.

9.3 Distance solutions

In Figure 9-7 we show a Cisco MDS 9509 Director configured with longwave SFPs and extended length products to create a distributed SAN. This solution demonstrates the maximum distances a SAN fabric can be separated by.

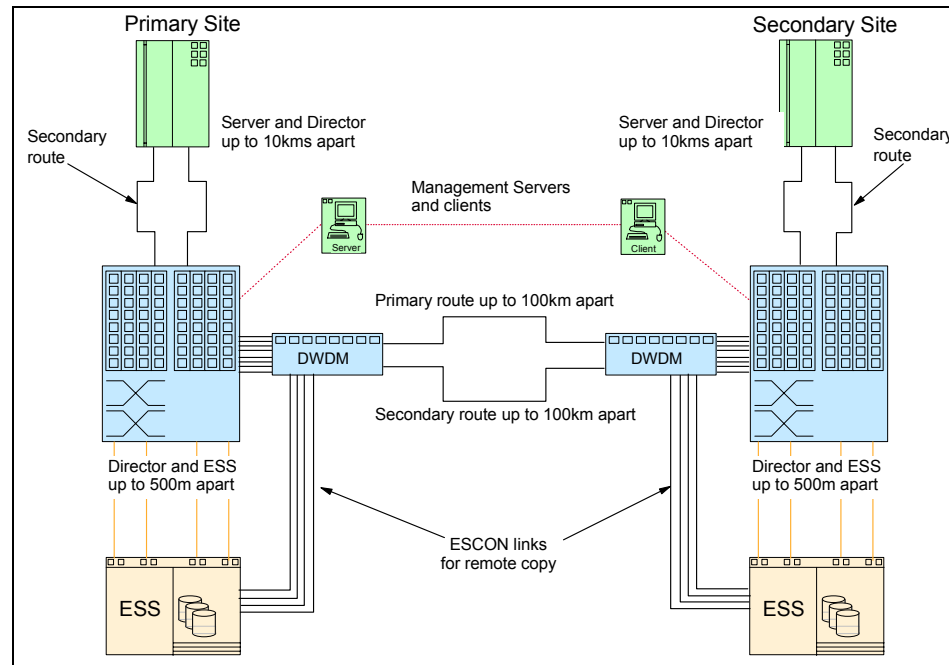


Figure 9-7 Solution using long distance SFPs

Fiber optic cable routes can either be leased from managed service providers, or private lines. If the lines are leased, it is important to consider data encryption as you may be sharing the same bandwidth with multiple companies. Lease line service providers normally provide reporting information, such as bandwidth utilization and link failures.

In our solution we have assumed we have our own private lines.

Due to the large distances between our fabric components and the increased risk of damage, wherever possible we have defined diverse cable routes. We have shown our diverse primary and secondary as different lengths as this is often the case if the lines are rented.

Our servers are using a HBA card that provides a single-mode SFP optic (9 micron). This HBA enables us to separate our servers from the director by up to 10 km.

We have used a Dense Wave Division Multiplexor (DWDM) to reduce our fiber optic cable line requirements between sites from twelve (4 x ESCON and 8 x FC links) to two while still retaining our original bandwidth.

Components

- ▶ SAN fabric:
 - 2 x Cisco MDS 9509 Directors configured with 64 ports (including two longwave ports for connection to the remote servers)
- ▶ Servers:
 - Servers each configured with dual FC HBAs with longwave capability
- ▶ Storage:
 - 2 x ESS-800 each configured with 8 x FC adapters and ESCON adapters for PPRC
 - PPRC license
- ▶ Dual DWDM with multiple input channels to handle 8 x FC ports from each ESS and 4 x ESCON channels for a total of 12 input channels per DWDM.
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS

Checklist

We checked the following items:

- ▶ Primary and secondary routes are diverse.
- ▶ Distances between primary and secondary routes are similar.
- ▶ We have longwave HBAs on our servers.
- ▶ Any application performance implications of longer distances.
- ▶ For leased lines, encryption is activated if required.
- ▶ For leased lines, bandwidth utilization reports are available.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

The major performance consideration with a long distance solution is calculating the correct number of lines between sites. This number can only be determined by performing a detailed performance profile of the servers and storage that will be remote. Over-estimating the number of lines will increase costs dramatically, and under-sizing the number of lines will dramatically effect the performance of the SAN.

It is vital that detailed performance data is available prior to sizing the number of lines required.

Typically, latency will increase over long distances, a good rule of thumb is 5 microseconds per kilometer. For further details on calculating the latency over distance refer to 3.5, “Buffers” on page 106.

The default E_D_TOV and R_A_TOV values do not need to be modified for this distance.

Scalability

Additional servers and storage devices can be connected to this solution using either shortwave or longwave SFPs depending on where they need to be located. With the Cisco MDS 9509 we can support from 32 to 224 ports within a single director and if more ports are required we can interconnect directors using E_Ports (or TE_Ports for VSANs). We can also scale a multi-director solution to implement PortChanneling whereby we can aggregate the bandwidth of a number of ISLs to form up to a 32 Gb/s logical connection between directors.

Security

Our solution assumes we own the private lines between sites so encryption is not required.

For leased lines or managed services where lines are shared, encryption is normally an option available from the service provider.

“What If” failure scenarios

These are some theoretical assumptions:

- ▶ If the primary route is significantly further than the secondary, this may have performance implications and will need verification.
- ▶ If a normal HBA (based on a shortwave SFP) is used, you will only be able to locate servers 500 m from the director.

9.4 Clustering solutions

In Figure 9-8 we create a highly available SAN with two Microsoft Exchange NT servers that are linked together using Microsoft Exchange installed on MSCS.

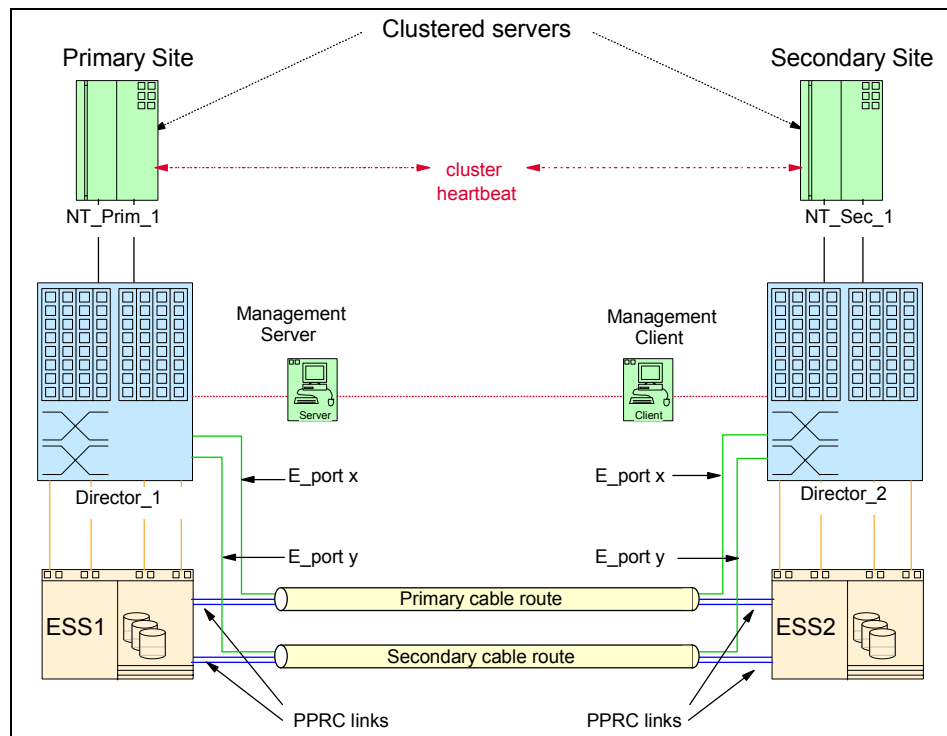


Figure 9-8 Clustered solution using Cisco MDS 9509 Director

Both NT_Prim_1 and NT_Sec_1 have dual HBAs. We are running dual pathing software, Subsystem Device Driver (SDD) to enable dynamic failover and workload balancing.

We are using 2 x 64 port Cisco MDS 9509 Directors which are connected using the E_Ports. The E_Ports x and y are both longwave SFPs and can be located on any two separate 16-port switching modules (all ports on a 16-port switching module support up to 255 buffer credits so we won't need to change this setting to support the increased distance).

We have defined the Cisco Fabric Manager and a remote client which is located at the secondary site. In the event of a site disaster the remote Fabric Manager client would be used to manage the fabric.

As we have only a single server and storage device there is no requirement to introduce any hard or name server zones, we could just use the default zone for this implementation. Zones (or even VSANs) could be required if we had a larger number of servers and we wanted to create SAN islands.

The director provides high availability so we have only configured one for each site. For additional resilience we could have two directors at each site linked using multiple E_Ports (or TE_Ports if we defined a VSAN).

To prevent connectivity issues due to an ESS bay being unavailable we have defined a FC connection into separate bays in the ESS. The ESSs are linked together using Peer to Peer Remote Copy (PPRC). We would need a minimum of four ESCON links between ESS1 and ESS2. The number of required links would need to be calculated based on the performance profile of the server.

The ESS requires LUN masking to be defined to enable servers to see their respective LUNs. We need to associate all NT_Prim_1 LUNs to its WWN, and we then must perform an ESS modify volume assignment and also define the WWN of the second HBA in NT_Prim_1 to the LUNs. We need to repeat this process for NT_Sec_1 WWNs. We now have 4 WWNs all allocated to the same LUNs on ESS1.

We must repeat this LUN masking process for ESS2, so ESS2 also has 4 WWNs defined. To support MS Clustering, the ESS microcode must be G3 or higher.

If we wanted to use NT_Prim_1 and NT_Sec_1 both as 'live' servers (that is to say, users connected and updating both servers) we could utilize the bi-directional function of PPRC. Changes made by NT_Prim_1 attached to ESS1 would be copied to ESS2; any changes made by NT_Sec_1 to ESS2 data would also be copied across ESS1.

We have diverse primary and secondary fibre cable routes to protect us from cable damage. If one route was to become unavailable, the secondary route would be used without loss of service.

Our solution does not include any extenders or repeater products, and therefore would be more suited to a campus solution. Without extenders or repeaters, our solution is limited to the maximum supported ESCON distance of 3 km. With extenders and repeaters we could increase these distances to 70 km.

Components

- ▶ SAN fabric:
 - 2 x Cisco MDS 9509 configured with 64 ports (each with two longwave ports)
- ▶ Servers:
 - 2 x servers configured with dual FC HBAs.
- ▶ Storage:
 - 2 x ESS-800's configured with 4 x Fibre Channel adapters and 4 x ESCON ports for Remote Copy in each machine.
 - PPRC license
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS
 - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- ▶ Version of cluster is supported with ESS.
- ▶ Version of cluster is supported with SDD.
- ▶ Applied recommended parameter changes to the HBAs to support clustering.
- ▶ Distance the cluster service is supported over.
- ▶ Cluster server failover time.
- ▶ Primary and secondary routes are diverse.
- ▶ All cables have been clearly labelled.
- ▶ Distances between primary and secondary routes are similar.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

In normal operation, the ISLs will only be used for replicating director configuration changes. Due to the FSPF principle the servers at both sites will only connect to their respective site director. There are no performance considerations for the ISL under normal operations.

It is good practice to monitor the load across the ISLs as other devices are added to the SAN. With this particular solution, there is very little chance of over-subscription of the ISLs, as the majority of the traffic flows over the local links. If we found that the ISLs were becoming a bottle neck, we could add an extra ISL, or we could look at using the PortChannel technology to create a larger capacity logical link between the directors. Using PortChannels would be a first step, as it would not require any additional infrastructure to implement, as it is setup using Cisco's Fabric Manager.

If the Cluster Service was activated, then NT_Sec_1 would automatically attach to ESS1 using the two E_Ports. In this scenario NT_Sec_1 would be reduced to 25% of its available SAN bandwidth connected to ESS1 as it would be using the ISLs to access the remote ESS.

We have a single director and the director is a full non-blocking device (as it is configured with multiple 16-port switching modules), there are no performance considerations for the director.

The important performance considerations for this solution are the number of fibers used between the director and ESS. In our example we have two server connections to two ESS connections. For more servers, this number will depend on the performance profile of the servers, and the potential performance impact of losing one or more switching modules.

In Section 9.1, "Performance solutions" on page 286 we look in more detail at performance considerations of having a larger number of servers connected using a director to a single ESS.

From an availability aspect we have configured a single shortwave connection from each ESS host bay to each switching module (4 connections). In the event of a single switching module failure we will have lost 25% of the available bandwidth to the local ESS.

Scalability

As with most of the solutions described in this section, there are many ways to scale the SAN. With this example we can add extra servers and storage devices into the existing infrastructure until we approach the limit of the number of ports available in the director.

With the blade architecture of the Cisco MDS 9000 family we can easily extend the existing infrastructure up to 224 ports. It's worth noting though that the Cisco MDS 9509 has 7-slots for switching modules which can be either 16 or 32 port modules. The 32-port switching modules should be used for lower performance devices that do not require the full 2 Gb/s throughput per port that is available with the 16-port switching module. The 32-port switching module can provide up to 32 Gb/s per switching module (or 4 Gb/s per port group).

The 16-port switching modules should be used for ISLs and for the servers and devices requiring a high performance profile.

When increasing the number of devices attached to the SAN we must also check that the ISLs are able to handle any increased load. We also need to consider the impact of a failover scenario where a server maybe accessing the storage remotely as this could cause contention at the ISL level.

Security

When operating with primary and secondary sites, we need to ensure that all related SAN documentation is in a secure location that can be accessed at the recovery site. We also need to ensure that we have enough userids of the correct type that are able to make any required changes to the zones.

“What If” failure scenarios

These are some theoretical assumptions:

- ▶ **Server:** If NT_Prim_1 fails the Microsoft cluster service would be activated and NT_Sec_1 will run NT_Prim_1's workload; obviously the server would need to be configured to handle this increased workload in the event of a cluster takeover.
- ▶ **Server HBA:** If an HBA fails SDD will automatically failover workload to the alternate HBA, dependent on application this could result in 30-40% performance degradation.
- ▶ **Cable:** If a cable between the server and director fails, SDD will automatically failover workload, and dependent on application this could result in 30-40% performance degradation.
- ▶ **Cable:** If a local cable fails between director and an ESS, an alternate route will be used, and we would lose up to 25% of the available bandwidth.
- ▶ **Cable:** If a cable fails between E_Ports there would be no effect in normal operation. In a cluster failover situation we would lose 50% of the available bandwidth.
- ▶ **Director:** If a switching module fails we still have connectivity as we have dual connections, but we would lose up to 50% bandwidth to any connected servers, and 25% bandwidth to the ESS.

- ▶ **Blade:** If a supervisor module fails there would be no effect as the spare supervisor module would automatically take over.
- ▶ **Backplane:** If the backplane was damaged we would lose connectivity to all servers at that site. Clustered servers would fail over to the secondary site.
- ▶ **ESS:** If an ESS HBA is unavailable, we have multiple other connections that will automatically be used. We would lose 25% of the available SAN bandwidth per HBA.
- ▶ **Bay:** If an ESS bay is unavailable we have multiple connections in other bays that will automatically be used. We would lose 25% of the available bandwidth per bay.
- ▶ **ESS:** If the ESS is unavailable we would automatically failover to ESS2.
- ▶ **SAN fabric:** If a problem occurs with the SAN fabric, then this could affect all servers in both the Primary and Secondary Sites. It represents a single point of failure.
- ▶ **Dual fabrics:** Our solution could be expanded to include dual fabrics to avoid the problem mentioned above, through the use of Cisco's VSAN technology. For further details of such a solution, refer to Section 9.2, "Availability solutions" on page 290.

9.5 Secure solutions

Any adverse effect to a SAN will typically have an impact on multiple servers within the SAN fabric. To minimize this impact, it is important to ensure that every possible security measure is incorporated into the SAN design.

9.5.1 Zoning security solution

In Figure 9-9 we look at a SAN design that could be used to provide additional security to the SAN. Using the Cisco VSAN technology we have created two isolated VSANs:

- ▶ VSAN_1 contains:
 - Our UNIX servers (Prod_UNIX and Devl_UNIX)
- ▶ VSAN_2 contains:
 - Our NT environment (to keep things simple, we have shown one NT server, but this could be multiple servers)

To further protect the servers, within each of the VSANs we have created three zones:

- ▶ Zone_1 contains the Production UNIX environment
- ▶ Zone_2 contains the Development UNIX environment
- ▶ Zone_3 contains the NT environment

By defining the zones within each VSAN we can further ensure that each server is protected and cannot access the other servers storage within the VSANs. For example, we don't want the Development server to ever be able to access the Production servers live data.

We could have defined a separate VSAN for our Development UNIX server but in our example we wanted to share ESS1 with the Production system as we plan to use FlashCopy® to populate our test databases. With VSANs, we cannot have a port in more than one VSAN, and this would have restricted the number of physical ports into the ESS that we would have access to. This in turn could have affected our ESS performance if we only had one path available.

To prevent unauthorized access to our SAN fabric, all SAN components are contained in a physically secure environment. The management interfaces to the ESS and MDS 9000 Director are userid and password protected, and it is not possible to define or amend these userids or password directly at the product.

To prevent unauthorized cable changes, the director has been installed in a locking cabinet, and this will protect us against unauthorized cable changes. For added protection against cable changes, we have also incorporated name server zoning. This way, if someone does move cables, the WWN will move with the cables and the zoning will still be accurate.

In this solution we have used the WWNs of the servers and created a name server zone through director 1 to the WWPN of the HBAs in each of the ESS. This compliments the LUN masking facility provided by the ESS.

The Cisco MDS 9000 family of products also allows for management authorization via a RADIUS server. In this example, we have utilized our existing RADIUS servers to authenticate any management request to update the MDS 9000 configuration. The following describes the authentication process:

- ▶ The user requests permission to perform a management task on the director.
- ▶ The director passes the request to the RADIUS server which checks the User Name, Password and User Access Roles.
- ▶ The RADIUS server accepts or denies the access request.
- ▶ The authentication process is either accepted or denied to the user.

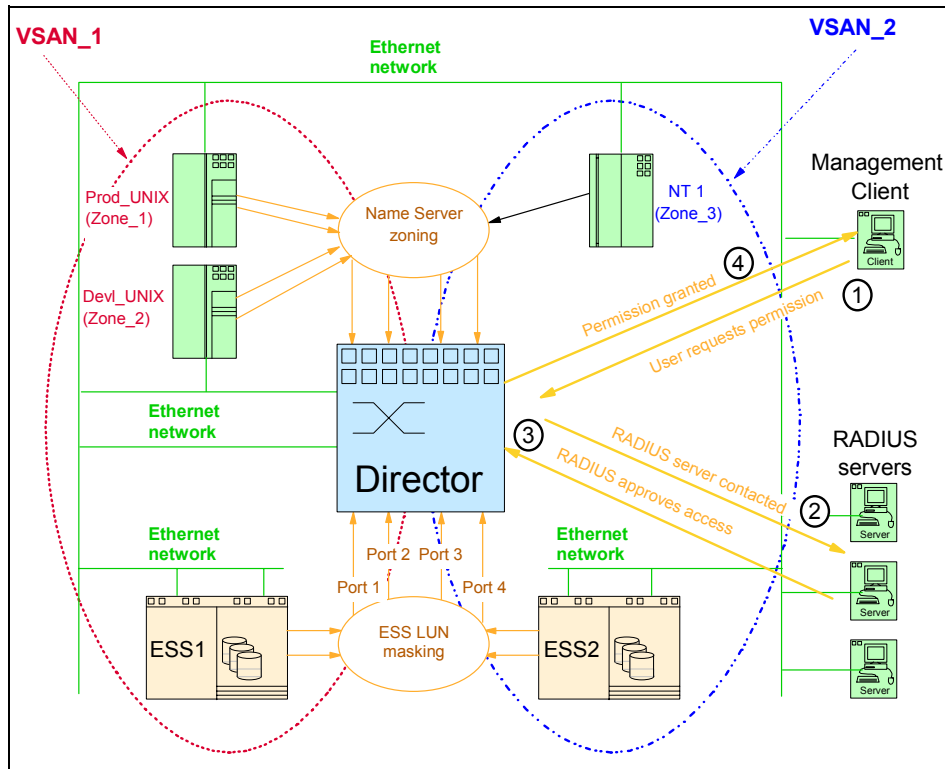


Figure 9-9 Protecting your data from both human error and sabotage

Components

- ▶ SAN fabric:
 - 1 x Cisco MDS 9509 configured with 64 ports
- ▶ Servers:
 - 2 x UNIX servers configured with dual FC HBAs
 - 1 x NT server
 - RADIUS servers
- ▶ Storage:
 - 2 x ESS-800's configured with 2 x FC adapters
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS
 - IOMETER and Disk Magic for performance modelling

Checklist

We checked the following items:

- ▶ All fabric components are in a secure location.
- ▶ The director is installed in a lockable cabinet in a secured location.
- ▶ The WWNs of the HBAs port names of NT1, Prod_UNIX and Prod_UNIX have been defined to a name server zone.
- ▶ If dual HBAs have been used, the LUNs defined on ESS1 and ESS2 have been associated with both HBA WWNs of NT1, Prod_UNIX and DevI_UNIX.
- ▶ The userids and passwords of the management tools have been changed from the default.
- ▶ Only SAN administrators have access to the Cisco Fabric Manager userids and passwords.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

This solution is looking at SAN security and therefore performance has not been considered in this case.

Scalability

We can currently define up to 1024 Virtual SANs within this fabric (VSAN 1 is a default VSAN and VSAN 4094 is the isolated VSAN leaving us up to 1022 available VSANs).

As mentioned previously, we can configured more hosts and storage devices by scaling up the number of ports within the director. We can also grow the director by connecting to multiple directors using E_Ports (or TE_Ports should we want to trunk VSANs across the ISLs).

Security

Security implications have been covered in the solution and checklist sections.

“What If” failure scenarios

These are some theoretical assumptions:

- ▶ Access is obtained to director cabinet. If the fibre optic cables belonging to NT1 Prod_UNIX or DevI_UNIX are switched, all the correct LUNs would still be visible at the server. Name server zoning is independent of the director port position.

- ▶ Access is obtained to the Cisco Fabric Manager server. Authentication process protects the fabric from an un-authorized user.
- ▶ Access is obtained to ESS Specialist. Even if ESS LUN masking information is modified and Prod_UNIX's HBA WWN is added to NT1's LUN list, it would not be possible for UNIX1 to see NT1's LUNs as it is in a separate VSAN. The ports for ESS2 are defined to VSAN_2 and are not accessible to VSAN_1.
- ▶ If someone managed to obtain access to the equipment and install a new device into the SAN, then these ports would be part of the default VSAN and would still not be accessible to either VSAN_1 or VSAN_2.

9.6 Loop solutions

The Cisco MDS 9000 family supports a wide range of different port devices without the requirement of additional fabric components. In the solution that follows, we illustrate how the Cisco MDS 9000 family can handle FC_AL connections using the Translative Loop (TL_Port).

9.6.1 Utilizing the translative loop port

In Figure 9-10 we are able to create a single SAN fabric that incorporates a mix of devices without having to allow for the “bad habits” of a loop device, particularly a LIP, or add any additional bridging devices.

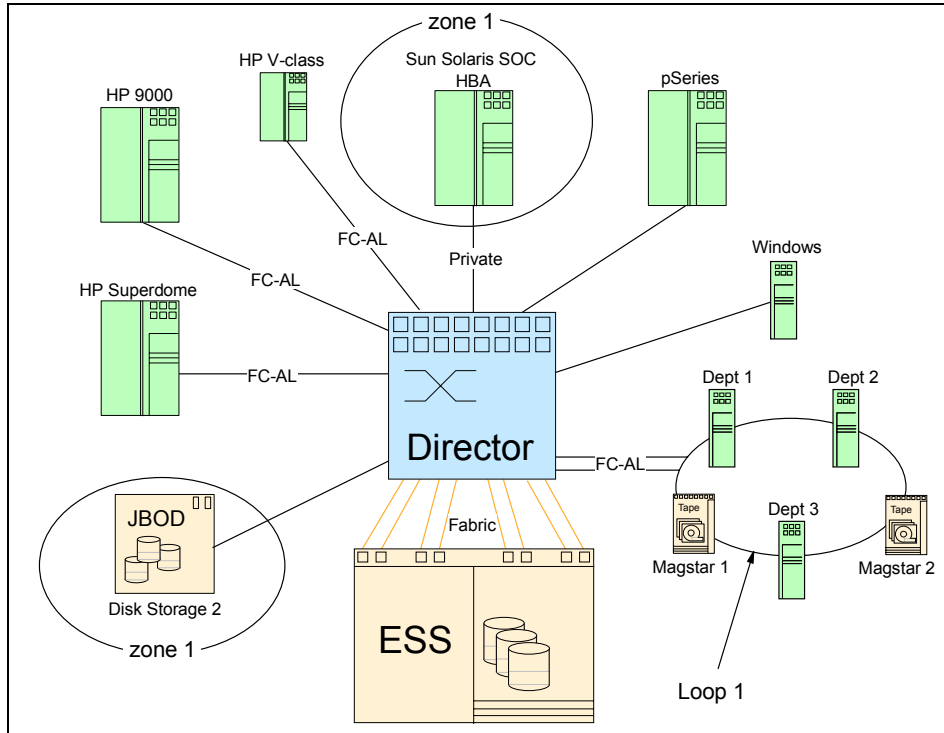


Figure 9-10 Utilizing the Cisco MDS 9000 TL_Port

For example, if Loop 1 was implemented as a traditional loop, that is to say, not connected to the Cisco MDS 9509, an FC-AL device error to Magstar 1 (which could be something as simple as someone pulling a tape from a tape library), will result in the departmental servers 1, 2, and 3, and Magstar 2 devices to be reset. This is due to the LIP which is sent across the loop to all devices.

If the Departmental Servers were performing I/O, or Magstar 2 was performing a backup, this LIP would provide an unwelcome interruption to processing.

In our example, Loop 1 is connected to a TL_Port of the director. As the TL_Port is able to treat Magstar 1 as a private device (even though it is connected to a public loop), an error to Magstar 1 will be transparent to the departmental servers and Magstar 2, as each of our loop devices are treated as individuals.

As we have the ability to tune the ports on the director we are able to connect our Sun server, which has an Adaptec JNI HBA installed and attach it to Disk Storage 2 using the Cisco MDS 9509 Director. The Adaptec HBA's are not supported by IBM for connection to an ESS so we must introduce zoning (soft zone 1) to ensure that only the Sun server is able to access this disk.

We are also able to attach all our HP servers (which only support FC-AL in this example), our IBM pSeries and Windows servers to the same SAN fabric, with no other additional SAN devices.

Components

- ▶ SAN fabric:
 - Cisco MDS 9509 Director configured with 4 x 16-port switching modules
- ▶ Servers:
 - 3 x HP servers each configured with dual FC HBAs
 - pSeries configured with dual HBAs
 - 4 x NT servers configured with single HBAs
- ▶ Storage:
 - ESS-800 configured with 8 x FC adapters
 - JBOD disk array with dual FC HBAs
- ▶ Software:
 - Cisco Fabric Manager
 - SDD installed on servers
 - ESS Specialist for configuring the ESS

Checklist

We checked the following items:

- ▶ Any servers that are not supported by IBM are connected to non-IBM disk.
- ▶ Any disk devices that do not support LUN masking are zoned to their respective servers.
- ▶ All storage devices, server HBAs, switches, or directors are configured with the latest supported versions of drivers/firmware levels.

Performance

For this solution we have no knowledge of our server's performance profile, so we assume our high performance profile of 6 server connections to each ESS storage port (refer to Section 9.1, "Performance solutions" on page 286 for details). For the Departmental Servers, we are using our low performance profile ratio of 12 server connections to each ESS Storage port. As the Magstar devices are part of the departmental loop, we will class the Magstar devices as low profile. We can ignore our Sun server, as this is attaching to its own device.

In our solution we have 5 high profile server connections to 1 ESS port and another 5 low profile device connections to one ESS port.

We are within our performance profile for both server groups.

We could also make use of the 32-port switching module by attaching servers that fit the low performance profile to this module. This will enable us to provide a better cost/port ratio and increase the port density while still managing the performance requirements of these servers.

Scalability

This solution can easily be scaled through the addition of more switching modules (either 16 or 32-port modules) and through the addition of extra switches connected via ISLs.

As there are no restrictions on how many TL_Ports can be defined or where they have to be located, we can easily build a solution to scale up to many hundreds of devices.

Security

The ESS performs LUN masking by default so all devices with LUNs defined have two levels of security: LUN masking and zoning. As Disk Storage 2 does not perform LUN masking, our only level of protection is zone 1. To guard against unauthorized changes to this zone, we could use the security management tools provided with Cisco's Fabric Manager — refer to Section 9.5, "Secure solutions" on page 308 for more details.

"What If" failure scenarios

These are some theoretical assumptions:

- ▶ If unauthorized access is obtained to the Cisco Fabric Manager, it would be possible to create a zone allowing any server to have access to Disk Storage 2's LUNs.
- ▶ If Magstar 2 is removed from Loop 1, there will be no effect on any of the other devices in Loop 1.
- ▶ The SUN server is connected to the ESS. This invalidates IBM's support of the SAN (the Sun server is configured with an HBA card that is not supported by IBM in this scenario), as you may not be able to see the LUNs belonging to the SUN server.
- ▶ One of the HBA ports on the ESS was defined as FC-AL, even though we are connecting through the director: The director will translate and allow the server to see the LUN.

Glossary

8b/10b A data encoding scheme developed by IBM, translating byte-wide data to an encoded 10-bit format. Fibre Channel's FC-1 level defines this as the method to be used to encode and decode data transmissions over the Fibre Channel.

Adapter A hardware unit that aggregates other I/O units, devices or communications links to a system bus.

ADSM ADSTAR Distributed Storage Manager.

Agent (1) In the client-server model, the part of the system that performs information preparation and exchange on behalf of a client or server application. (2) In SNMP, the word agent refers to the managed system. See also: Management Agent.

Aggregation In the Storage Networking Industry Association Storage Model (SNIA), *virtualization* is known as *aggregation*. This aggregation can take place at the file level or at the level of individual blocks that are transferred to disk.

AIT Advanced Intelligent Tape - A magnetic tape format by Sony that uses 8mm cassettes, but is only used in specific drives.

AL See Arbitrated Loop

AL_PA Arbitrated Loop Physical Address

ANSI American National Standards Institute - The primary organization for fostering the development of technology standards in the United States. The ANSI family of Fibre Channel documents provide the standards

basis for the Fibre Channel architecture and technology. See FC-PH

Arbitration The process of selecting one respondent from a collection of several candidates that request service concurrently.

Arbitrated Loop A Fibre Channel interconnection technology that allows up to 126 participating node ports and one participating fabric port to communicate.

ATL Automated Tape Library - Large scale tape storage system, which uses multiple tape drives and mechanisms to address 50 or more cassettes.

ATM Asynchronous Transfer Mode - A type of packet switching that transmits fixed-length units of data.

Backup A copy of computer data that is used to recreate data that has been lost, mislaid, corrupted, or erased. The act of creating a copy of computer data that can be used to recreate data that has been lost, mislaid, corrupted or erased.

Bandwidth Measure of the information capacity of a transmission channel.

Bridge (1) A component used to attach more than one I/O unit to a port. (2) A data communications device that connects two or more networks and forwards packets between them. The bridge may use similar or dissimilar media and signaling systems. It operates at the data link level of the OSI model. Bridges read and filter data packets and frames.

Bridge/Router A device that can provide the functions of a bridge, router or both concurrently. A bridge/router can route one or more protocols, such as TCP/IP, and bridge all other traffic. See also: Bridge, Router.

Broadcast Sending a transmission to all N_Ports on a fabric.

Channel A point-to-point link, the main task of which is to transport data from one point to another.

Channel I/O A form of I/O where request and response correlation is maintained through some form of source, destination and request identification.

CIFS Common Internet File System

Class of Service A Fibre Channel frame delivery scheme exhibiting a specified set of delivery characteristics and attributes.

Class-1 A class of service providing dedicated connection between two ports with confirmed delivery or notification of non-deliverability.

Class-2 A class of service providing a frame switching service between two ports with confirmed delivery or notification of non-deliverability.

Class-3 A class of service providing frame switching datagram service between two ports or a multicast service between a multicast originator and one or more multicast recipients.

Class-4 A class of service providing a fractional bandwidth virtual circuit between two ports with confirmed delivery or notification of non-deliverability.

Class-6 A class of service providing a multicast connection between a multicast

originator and one or more multicast recipients with confirmed delivery or notification of non-deliverability.

Client A software program used to contact and obtain data from a *server* software program on another computer -- often across a great distance. Each *client* program is designed to work specifically with one or more kinds of server programs and each server requires a specific kind of client program.

Client/Server The relationship between machines in a communications network. The client is the requesting machine, the server the supplying machine. Also used to describe the information management relationship between software components in a processing system.

Cluster A type of parallel or distributed system that consists of a collection of interconnected whole computers and is used as a single, unified **computing resource**.

Coaxial Cable A transmission media (cable) used for high speed transmission. It is called *coaxial* because it includes one physical channel that carries the signal surrounded (after a layer of insulation) by another concentric physical channel, both of which run along the same axis. The inner channel carries the signal and the outer channel serves as a ground.

Controller A component that attaches to the system topology through a channel semantic protocol that includes some form of request/response identification.

CRC Cyclic Redundancy Check - An error-correcting code used in Fibre Channel.

DASD Direct Access Storage Device - any on-line storage device: a disc, drive or CD-ROM.

DAT Digital Audio Tape - A tape media technology designed for very high quality audio recording and data backup. DAT cartridges look like audio cassettes and are often used in mechanical auto-loaders. typically, a DAT cartridge provides 2GB of storage. But new DAT systems have much larger capacities.

Data Sharing A SAN solution in which files on a storage device are shared between multiple hosts.

Datagram Refers to the Class 3 Fibre Channel Service that allows data to be sent rapidly to multiple devices attached to the fabric, with no confirmation of delivery.

dB Decibel - a ratio measurement distinguishing the percentage of signal attenuation between the input and output power. Attenuation (loss) is expressed as dB/km

Disk Mirroring A fault-tolerant technique that writes data simultaneously to two hard disks using the same hard disk controller.

Disk Pooling A SAN solution in which disk storage resources are pooled across multiple hosts rather than be dedicated to a specific host.

DLT Digital Linear Tape - A magnetic tape technology originally developed by Digital Equipment Corporation (DEC) and now sold by Quantum. DLT cartridges provide storage capacities from 10 to 35GB.

E_Port Expansion Port - a port on a switch used to link multiple switches together into a Fibre Channel switch fabric.

ECL Emitter Coupled Logic - The type of transmitter used to drive copper media such as Twinax, Shielded Twisted Pair, or Coax.

Enterprise Network A geographically dispersed network under the auspices of one organization.

Entity In general, a real or existing thing from the Latin ens, or being, which makes the distinction between a thing's existence and its qualities. In programming, engineering and probably many other contexts, the word is used to identify units, whether concrete things or abstract ideas, that have no ready name or label.

ESCON Enterprise System Connection

Exchange A group of sequences which share a unique identifier. All sequences within a given exchange use the same protocol. Frames from multiple sequences can be multiplexed to prevent a single exchange from consuming all the bandwidth. See also: Sequence

F_Node Fabric Node - a fabric attached node.

F_Port Fabric Port - a port used to attach a Node Port (N_Port) to a switch fabric.

Fabric Fibre Channel employs a fabric to connect devices. A fabric can be as simple as a single cable connecting two devices. The term is most often used to describe a more complex network utilizing hubs, switches and gateways.

Fabric Login Fabric Login (FLOGI) is used by an N_Port to determine if a fabric is present and, if so, to initiate a session with the fabric by exchanging service parameters with the fabric. Fabric Login is performed by an N_Port following link initialization and before communication with other N_Ports is attempted.

FC Fibre Channel

FC-0 Lowest level of the Fibre Channel Physical standard, covering the physical characteristics of the interface and media

FC-1 Middle level of the Fibre Channel Physical standard, defining the 8b/10b encoding/decoding and transmission protocol.

FC-2 Highest level of the Fibre Channel Physical standard, defining the rules for signaling protocol and describing transfer of frame, sequence and exchanges.

FC-3 The hierarchical level in the Fibre Channel standard that provides common services such as striping definition.

FC-4 The hierarchical level in the Fibre Channel standard that specifies the mapping of upper-layer protocols to levels below.

FCA Fibre Channel Association.

FC-AL Fibre Channel Arbitrated Loop - A reference to the Fibre Channel Arbitrated Loop standard, a shared gigabit media for up to 127 nodes, one of which may be attached to a switch fabric. See also: Arbitrated Loop.

FC-CT Fibre Channel common transport protocol

FC-FG Fibre Channel Fabric Generic - A reference to the document (ANSI X3.289-1996) which defines the concepts, behavior and characteristics of the Fibre Channel Fabric along with suggested partitioning of the 24-bit address space to facilitate the routing of frames.

FC-FP Fibre Channel HIPPI Framing Protocol - A reference to the document (ANSI X3.254-1994) defining how the HIPPI framing protocol is transported via the Fibre Channel

FC-GS Fibre Channel Generic Services - A reference to the document (ANSI X3.289-1996) describing a common transport protocol used to communicate with the server functions, a full X500 based directory service, mapping of the Simple Network Management Protocol (SNMP) directly to the Fibre Channel, a time server and an alias server.

FC-LE Fibre Channel Link Encapsulation - A reference to the document (ANSI X3.287-1996) which defines how IEEE 802.2 Logical Link Control (LLC) information is transported via the Fibre Channel.

FC-PH A reference to the Fibre Channel Physical and Signaling standard ANSI X3.230, containing the definition of the three lower levels (FC-0, FC-1, and FC-2) of the Fibre Channel.

FC-PLDA Fibre Channel Private Loop Direct Attach - See PLDA.

FC-SB Fibre Channel Single Byte Command Code Set - A reference to the document (ANSI X.271-1996) which defines how the ESCON command set protocol is transported using the Fibre Channel.

FC-SW Fibre Channel Switch Fabric - A reference to the ANSI standard under development that further defines the fabric behavior described in FC-FG and defines the communications between different fabric elements required for those elements to coordinate their operations and management address assignment.

FC Storage Director See SAN Storage Director

FCA Fibre Channel Association - a Fibre Channel industry association that works to promote awareness and understanding of the Fibre Channel technology and its application

and provides a means for implementers to support the standards committee activities.

FCLC Fibre Channel Loop Association - an independent working group of the Fibre Channel Association focused on the marketing aspects of the Fibre Channel Loop technology.

FCP Fibre Channel Protocol - the mapping of SCSI-3 operations to Fibre Channel.

Fiber Optic Refers to the medium and the technology associated with the transmission of information along a glass or plastic wire or fiber.

Fibre Channel A technology for transmitting data between computer devices at a data rate of up to 4 Gb/s. It is especially suited for connecting computer servers to shared storage devices and for interconnecting storage controllers and drives.

FICON Fibre Connection - A next-generation I/O solution for IBM S/390 parallel enterprise server.

FL_Port Fabric Loop Port - the access point of the fabric for physically connecting the user's Node Loop Port (NL_Port).

FLOGI See Fabric Log In

Frame A linear set of transmitted bits that define the basic transport unit. The frame is the most basic element of a message in Fibre Channel communications, consisting of a 24-byte header and zero to 2112 bytes of data. See also: Sequence

FSP Fibre Channel Service Protocol - The common FC-4 level protocol for all services, transparent to the fabric type or topology.

FSPF Fabric Shortest Path First - is an intelligent path selection and routing standard and is part of the Fibre Channel Protocol.

Full-Duplex A mode of communications allowing simultaneous transmission and reception of frames.

G_Port Generic Port - a generic switch port that is either a Fabric Port (F_Port) or an Expansion Port (E_Port). The function is automatically determined during login.

Gateway A node on a network that interconnects two otherwise incompatible networks.

Gb/s Gigabits per second. Also sometimes referred to as Gbps. In computing terms it is approximately 1,000,000,000 bits per second. Most precisely it is 1,073,741,824 (1024 x 1024 x 1024) bits per second.

GB/s Gigabytes per second. Also sometimes referred to as GBps. In computing terms it is approximately 1,000,000,000 bytes per second. Most precisely it is 1,073,741,824 (1024 x 1024 x 1024) bytes per second.

GBIC GigaBit Interface Converter - Industry standard transceivers for connection of Fibre Channel nodes to arbitrated loop hubs and fabric switches.

Gigabit One billion bits, or one thousand megabits.

GLM Gigabit Link Module - a generic Fibre Channel transceiver unit that integrates the key functions necessary for installation of a Fibre Channel media interface on most systems.

Half-Duplex A mode of communications allowing either transmission or reception of frames at any point in time, but not both (other

than link control frames which are always permitted).

Hardware The mechanical, magnetic and electronic components of a system, for example, computers, telephone switches, terminals and the like.

HBA Host Bus Adapter

HIPPI High Performance Parallel Interface - An ANSI standard defining a channel that transfers data between CPUs and from a CPU to disk arrays and other peripherals.

HMMP HyperMedia Management Protocol

HMMS HyperMedia Management Schema - the definition of an implementation-independent, extensible, common data description/schema allowing data from a variety of sources to be described and accessed in real time regardless of the source of the data. See also: WBEM, HMMP

hop A FC frame may travel from a switch to a director, a switch to a switch, or director to a director which, in this case, is one hop.

HSM Hierarchical Storage Management - A software and hardware system that moves files from disk to slower, less expensive storage media based on rules and observation of file activity. Modern HSM systems move files from magnetic disk to optical disk to magnetic tape.

HUB A Fibre Channel device that connects nodes into a logical loop by using a physical star topology. Hubs will automatically recognize an active node and insert the node into the loop. A node that fails or is powered off is automatically removed from the loop.

HUB Topology see Loop Topology

Hunt Group A set of associated Node Ports (N_Ports) attached to a single node, assigned a special identifier that allows any frames containing this identifier to be routed to any available Node Port (N_Port) in the set.

In-band Signaling This is signaling that is carried in the same channel as the information. Also referred to as in-band.

In-band virtualization An implementation in which the virtualization process takes place in the data path between servers and disk systems. The virtualization can be implemented as software running on servers or in dedicated engines.

Information Unit A unit of information defined by an FC-4 mapping. Information Units are transferred as a Fibre Channel Sequence.

Intermix A mode of service defined by Fibre Channel that reserves the full Fibre Channel bandwidth for a dedicated Class 1 connection, but also allows connection-less Class 2 traffic to share the link if the bandwidth is available.

Inter switch link A FC connection between switches and/or directors. Also known as ISL.

I/O Input/output

IP Internet Protocol

IPI Intelligent Peripheral Interface

ISL See Inter switch link.

Isochronous Transmission Data transmission which supports network-wide timing requirements. A typical application for isochronous transmission is a broadcast environment which needs information to be delivered at a predictable time.

JBOD Just a bunch of disks.

Jukebox A device that holds multiple optical disks and one or more disk drives, and can swap disks in and out of the drive as needed.

L_Port Loop Port - A node or fabric port capable of performing Arbitrated Loop functions and protocols. NL_Ports and FL_Ports are loop-capable ports.

LAN See Local Area Network - A network covering a relatively small geographic area (usually not larger than a floor or small building). Transmissions within a Local Area Network are mostly digital, carrying data among stations at rates usually above one megabit/s.

Latency A measurement of the time it takes to send a frame between two locations.

LC Lucent Connector. A registered trademark of Lucent Technologies.

Link A connection between two Fibre Channel ports consisting of a transmit fibre and a receive fibre.

Link_Control_Facility A termination card that handles the logical and physical control of the Fibre Channel link for each mode of use.

LIP A Loop Initialization Primitive sequence is a special Fibre Channel sequence that is used to start loop initialization. Allows ports to establish their port addresses.

Local Area Network (LAN) A network covering a relatively small geographic area (usually not larger than a floor or small building). Transmissions within a Local Area Network are mostly digital, carrying data among stations at rates usually above one megabit/s.

Login Server Entity within the Fibre Channel fabric that receives and responds to login requests.

Loop Circuit A temporary point-to-point like path that allows bi-directional communications between loop-capable ports.

Loop Topology An interconnection structure in which each point has physical links to two neighbors resulting in a closed circuit. In a loop topology, the available bandwidth is shared.

LVD Low Voltage Differential

Management Agent A process that exchanges a managed node's information with a management station.

Managed Node A managed node is a computer, a storage system, a gateway, a media device such as a switch or hub, a control instrument, a software product such as an operating system or an accounting package, or a machine on a factory floor, such as a robot.

Managed Object A variable of a managed node. This variable contains one piece of information about the node. Each node can have several objects.

Management Station A host system that runs the management software.

MAR Media Access Rules. Enable systems to self-configure themselves is a SAN environment

Mb/s Megabits per second. Also sometimes referred to as Mbps. In computing terms it is approximately 1,000,000 bits per second. Most precisely it is 1,048,576 (1024 x 1024) bits per second.

MB/s Megabytes per second. Also sometimes referred to as MBps. In computing terms it is approximately 1,000,000 bytes per second. Most precisely it is 1,048,576 (1024 x 1024) bytes per second.

Metadata server In Storage Tank, servers that maintain information (metadata) about the data files and grant permission for application servers to communicate directly with disk systems.

Meter 39.37 inches, or just slightly larger than a yard (36 inches)

Media Plural of medium. The physical environment through which transmission signals pass. Common media include copper and fiber optic cable.

Media Access Rules (MAR).

MIA Media Interface Adapter - MIAs enable optic-based adapters to interface to copper-based devices, including adapters, hubs, and switches.

MIB Management Information Block - A formal description of a set of network objects that can be managed using the Simple Network Management Protocol (SNMP). The format of the MIB is defined as part of SNMP and is a hierarchical structure of information relevant to a specific device, defined in object oriented terminology as a collection of objects, relations, and operations among objects.

Mirroring The process of writing data to two separate physical devices simultaneously.

MM Multi-Mode - See Multi-Mode Fiber

MMF See Multi-Mode Fiber - - In optical fiber technology, an optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle

within the optical core. Multi-Mode fiber transmission is used for relatively short distances because the modes tend to disperse over longer distances. See also: Single-Mode Fiber, SMF

Multicast Sending a copy of the same transmission from a single source device to multiple destination devices on a fabric. This includes sending to all N_Ports on a fabric (broadcast) or to only a subset of the N_Ports on a fabric (multicast).

Multi-Mode Fiber (MMF) In optical fiber technology, an optical fiber that is designed to carry multiple light rays or modes concurrently, each at a slightly different reflection angle within the optical core. Multi-Mode fiber transmission is used for relatively short distances because the modes tend to disperse over longer distances. See also: Single-Mode Fiber

Multiplex The ability to intersperse data from multiple sources and destinations onto a single transmission medium. Refers to delivering a single transmission to multiple destination Node Ports (N_Ports).

N_Port Node Port - A Fibre Channel-defined hardware entity at the end of a link which provides the mechanisms necessary to transport information units to or from another node.

N_Port Login N_Port Login (PLOGI) allows two N_Ports to establish a session and exchange identities and service parameters. It is performed following completion of the fabric login process and prior to the FC-4 level operations with the destination port. N_Port Login may be either explicit or implicit.

Name Server Provides translation from a given node name to one or more associated N_Port identifiers.

NAS Network Attached Storage - a term used to describe a technology where an integrated storage system is attached to a messaging network that uses common communications protocols, such as TCP/IP.

NDMP Network Data Management Protocol

Network An aggregation of interconnected nodes, workstations, file servers, and/or peripherals, with its own protocol that supports interaction.

Network Topology Physical arrangement of nodes and interconnecting communications links in networks based on application requirements and geographical distribution of users.

NFS Network File System - A distributed file system in UNIX developed by Sun Microsystems which allows a set of computers to cooperatively access each other's files in a transparent manner.

NL_Port Node Loop Port - a node port that supports Arbitrated Loop devices.

NMS Network Management System - A system responsible for managing at least part of a network. NMSs communicate with agents to help keep track of network statistics and resources.

Node An entity with one or more N_Ports or NL_Ports.

Non-Blocking A term used to indicate that the capabilities of a switch are such that the total number of available transmission paths is equal to the number of ports. Therefore, all ports can have simultaneous access through the switch.

Non-L_Port A Node or Fabric port that is not capable of performing the Arbitrated Loop

functions and protocols. N_Ports and F_Ports are not loop-capable ports.

Operation A term defined in FC-2 that refers to one of the Fibre Channel *building blocks* composed of one or more, possibly concurrent, exchanges.

Optical Disk A storage device that is written and **read by laser light**.

Optical Fiber A medium and the technology associated with the transmission of information as light pulses along a glass or plastic wire or fiber.

Ordered Set A Fibre Channel term referring to four 10-bit characters (a combination of data and special characters) providing low-level link functions, such as frame demarcation and signaling between two ends of a link.

Originator A Fibre Channel term referring to the initiating device.

Out of Band Signaling This is signaling that is separated from the channel carrying the information. Also referred to as out-of-band.

Out-of-band virtualization An alternative type of virtualization in which servers communicate directly with disk systems under control of a virtualization function that is not involved in the data transfer.

Peripheral Any computer device that is not part of the essential computer (the processor, memory and data paths) but is situated relatively close by. A near synonym is input/output (I/O) device.

Petard A device that is small and sometimes explosive.

PLDA Private Loop Direct Attach - A technical report which defines a subset of the relevant

standards suitable for the operation of peripheral devices such as disks and tapes on a private loop.

PLOGI See N_Port Login

Point-to-Point Topology An interconnection structure in which each point has physical links to only one neighbor resulting in a closed circuit. In point-to-point topology, the available bandwidth is dedicated.

Policy-based management Management of data on the basis of business policies (for example, “all production database data must be backed up every day”), rather than technological considerations (for example, “all data stored on this disk system is protected by remote copy”).

Port The hardware entity within a node that performs data communications over the Fibre Channel.

Port Bypass Circuit A circuit used in hubs and disk enclosures to automatically open or close the loop to add or remove nodes on the loop.

Private NL_Port An NL_Port which does not attempt login with the fabric and only communicates with other NL Ports on the same loop.

Protocol A data transmission convention encompassing timing, control, formatting and data representation.

Public NL_Port An NL_Port that attempts login with the fabric and can observe the rules of either public or private loop behavior. A public NL_Port may communicate with both private and public NL_Ports.

Quality of Service (QoS) A set of communications characteristics required by an

application. Each QoS defines a specific transmission priority, level of route reliability, and security level.

Quick Loop is a unique Fibre Channel topology that combines arbitrated loop and fabric topologies. It is an optional licensed product that allows arbitrated loops with private devices to be attached to a fabric.

RAID Redundant Array of Inexpensive or Independent Disks. A method of configuring multiple disk drives in a storage subsystem for high availability and high performance.

Raid 0 Level 0 RAID support - Striping, no redundancy

Raid 1 Level 1 RAID support - mirroring, complete redundancy

Raid 5 Level 5 RAID support, Striping with parity

Repeater A device that receives a signal on an electromagnetic or optical transmission medium, amplifies the signal, and then retransmits it along the next leg of the medium.

Responder A Fibre Channel term referring to the answering device.

Router (1) A device that can decide which of several paths network traffic will follow based on some optimal metric. Routers forward packets from one network to another based on network-layer information. (2) A dedicated computer hardware and/or software package which manages the connection between two or more networks. See also: Bridge, Bridge/Router

SAF-TE SCSI Accessed Fault-Tolerant Enclosures

SAN A Storage Area Network (SAN) is a dedicated, centrally managed, secure information infrastructure, which enables any-to-any interconnection of servers and storage systems.

SAN System Area Network - term originally used to describe a particular symmetric multiprocessing (SMP) architecture in which a switched interconnect is used in place of a shared bus. Server Area Network - refers to a switched interconnect between multiple SMPs.

SANSymphony In-band block-level virtualization software made by DataCore Software Corporation and resold by IBM.

SC Connector A fiber optic connector standardized by ANSI TIA/EIA-568A for use in structured wiring installations.

Scalability The ability of a computer application or product (hardware or software) to continue to function well as it (or its context) is changed in size or volume. For example, the ability to retain performance levels when adding additional processors, memory and/or storage.

SCSI Small Computer System Interface - A set of evolving ANSI standard electronic interfaces that allow personal computers to communicate with peripheral hardware such as disk drives, tape drives, CD_ROM drives, printers and scanners faster and more flexibly than previous interfaces. The table below identifies the major characteristics of the different SCSI version.

SCSI Version	Signal Rate MHz	Bus Width (bits)	Max. DTR (MB/s)	Max. Num. Devices	Max. Cable Length (m)
SCSI-1	5	8	5	7	6
SCSI-2	5	8	5	7	6

Wide SCSI-2	5	16	10	15	6
Fast SCSI-2	10	8	10	7	6
Fast Wide SCSI-2	10	16	20	15	6
Ultra SCSI	20	8	20	7	1.5
Ultra SCSI-2	20	16	40	7	12
Ultra2 LVD SCSI	40	16	80	15	12

SCSI-3 SCSI-3 consists of a set of primary commands and additional specialized command sets to meet the needs of specific device types. The SCSI-3 command sets are used not only for the SCSI-3 parallel interface but for additional parallel and serial protocols, including Fibre Channel, Serial Bus Protocol (used with IEEE 1394 Firewire physical protocol) and the Serial Storage Protocol (SSP).

SCSI-FCP The term used to refer to the ANSI Fibre Channel Protocol for SCSI document (X3.269-199x) that describes the FC-4 protocol mappings and the definition of how the SCSI protocol and command set are transported using a Fibre Channel interface.

Sequence A series of frames strung together in numbered order which can be transmitted over a Fibre Channel connection as a single operation. See also: Exchange

SERDES Serializer Deserializer

Server A computer which is dedicated to one task.

SES SCSI Enclosure Services - ANSI SCSI-3 proposal that defines a command set for soliciting basic device status (temperature, fan

speed, power supply status, etc.) from a storage enclosures.

Single-Mode Fiber In optical fiber technology, an optical fiber that is designed for the transmission of a single ray or mode of light as a carrier. It is a single light path used for long-distance signal transmission. See also: Multi-Mode Fiber

SMART Self Monitoring and Reporting Technology

SM Single Mode - See Single-Mode Fiber

SMF Single-Mode Fiber - In optical fiber technology, an optical fiber that is designed for the transmission of a single ray or mode of light as a carrier. It is a single light path used for long-distance signal transmission. See also: MMF

SNIA Storage Networking Industry Association. A non-profit organization comprised of more than 77 companies and individuals in the storage industry.

SN Storage Network. See also: SAN

SNMP Simple Network Management Protocol - The Internet network management protocol which provides a means to monitor and set network configuration and run-time parameters.

SNMWG Storage Network Management Working Group is chartered to identify, define and support open standards needed to address the increased management requirements imposed by storage area network environments.

SSA Serial Storage Architecture - A high speed serial loop-based interface developed as a high speed point-to-point connection for

peripherals, particularly high speed storage arrays, RAID and CD-ROM storage by IBM.

Star The physical configuration used with hubs in which each user is connected by communications links radiating out of a central hub that handles all communications.

Storage Tank An IBM file aggregation project that enables a pool of storage, and even individual files, to be shared by servers of different types. In this way, Storage Tank can greatly improve storage utilization and enables data sharing.

StorWatch Expert These are StorWatch applications that employ a 3 tiered architecture that includes a management interface, a StorWatch manager and agents that run on the storage resource(s) being managed. Expert products employ a StorWatch data base that can be used for saving key management data (for example, capacity or performance metrics). Expert products use the agents as well as analysis of storage data saved in the data base to perform higher value functions including -- reporting of capacity, performance, etc. over time (trends), configuration of multiple devices based on policies, monitoring of capacity and performance, automated responses to events or conditions, and storage related data mining.

StorWatch Specialist A StorWatch interface for managing an individual fibre Channel device or a limited number of like devices (that can be viewed as a single group). StorWatch specialists typically provide simple, point-in-time management functions such as configuration, reporting on asset and status information, simple device and event monitoring, and perhaps some service utilities.

Striping A method for achieving higher bandwidth using multiple N_Ports in parallel to

transmit a single information unit across multiple levels.

STP Shielded Twisted Pair

Storage Media The physical device itself, onto which data is recorded. Magnetic tape, optical disks, floppy disks are all storage media.

Switch A component with multiple entry/exit points (ports) that provides dynamic connection between any two of these points.

Switch Topology An interconnection structure in which any entry point can be dynamically connected to any exit point. In a switch topology, the available bandwidth is scalable.

T11 A technical committee of the National Committee for Information Technology Standards, titled T11 I/O Interfaces. It is tasked with developing standards for moving data in and out of computers.

Tape Backup Making magnetic tape copies of hard disk and optical disc files for disaster recovery.

Tape Pooling A SAN solution in which tape resources are pooled and shared across multiple hosts rather than being dedicated to a specific host.

TCP Transmission Control Protocol - a reliable, full duplex, connection-oriented end-to-end transport protocol running on top of IP.

TCP/IP Transmission Control Protocol/ Internet Protocol - a set of communications protocols that support peer-to-peer connectivity functions for both local and wide area networks.

Time Server A Fibre Channel-defined service function that allows for the management of all timers used within a Fibre Channel system.

Topology An interconnection scheme that allows multiple Fibre Channel ports to communicate. For example, point-to-point, Arbitrated Loop, and switched fabric are all Fibre Channel topologies.

TL_Port A private to public bridging of switches or directors, referred to as Translative Loop.

Twinax A transmission media (cable) consisting of two insulated central conducting leads of coaxial cable.

Twisted Pair A transmission media (cable) consisting of two insulated copper wires twisted around each other to reduce the induction (thus interference) from one wire to another. The twists, or lays, are varied in length to reduce the potential for signal interference between pairs. Several sets of twisted pair wires may be enclosed in a single cable. This is the most common type of transmission media.

ULP Upper Level Protocols

UTC Under-The-Covers, a term used to characterize a subsystem in which a small number of hard drives are mounted inside a higher function unit. The power and cooling are obtained from the system unit. Connection is by parallel copper ribbon cable or pluggable backplane, using IDE or SCSI protocols.

UTP Unshielded Twisted Pair

Virtual Circuit A unidirectional path between two communicating N_Ports that permits fractional bandwidth.

Virtualization An abstraction of storage where the representation of a storage unit to the operating system and applications on a server is divorced from the actual physical storage where the information is contained.

Virtualization engine Dedicated hardware and software that is used to implement virtualization.

WAN Wide Area Network - A network which encompasses inter-connectivity between devices over a wide geographic area. A wide area network may be privately owned or rented, but the term usually connotes the inclusion of public (shared) networks.

WDM Wave Division Multiplexing - A technology that puts data from different sources together on an optical fiber, with each signal carried on its own separate light wavelength. Using WDM, up to 80 (and theoretically more) separate wavelengths or channels of data can be multiplexed into a stream of light transmitted on a single optical fiber.

WBEM Web-Based Enterprise Management - A consortium working on the development of a series of standards to enable active management and monitoring of network-based elements.

Zoning In Fibre Channel environments, the grouping together of multiple ports to form a virtual private storage network. Ports that are members of a group or zone can communicate with each other but are isolated from ports in other zones.

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

IBM Redbooks

- ▶ *Designing and Optimizing an IBM Storage Area Network*, SG24-6419
- ▶ *Designing an IBM Storage Area Network*, SG24-5758
- ▶ *Introduction to SAN Distance Solutions*, SG24-6408
- ▶ *Introducing Hosts to the SAN fabric*, SG24-6411
- ▶ *Implementing an Open IBM SAN*, SG24-6116
- ▶ *Introduction to Storage Area Networks*, SG24-5470
- ▶ *IP Storage Networking: IBM NAS and iSCSI Solutions*, SG24-6240
- ▶ *The IBM TotalStorage NAS 200 and 300 Integration Guide*, SG24-6505
- ▶ *Implementing the IBM TotalStorage NAS 300G: High Speed Cross Platform Storage and Tivoli SANergy!*, SG24-6278
- ▶ *iSCSI Performance Testing & Tuning*, SG24-6531
- ▶ *Using iSCSI Solutions' Planning and Implementation*, SG24-6291
- ▶ *Storage Networking Virtualization: What's it all about?*, SG24-6210
- ▶ *IBM Storage Solutions for Server Consolidation*, SG24-5355
- ▶ *Implementing the Enterprise Storage Server in Your Environment*, SG24-5420
- ▶ *Implementing Linux with IBM Disk Storage*, SG24-6261
- ▶ *Storage Area Networks: Tape Future In Fabrics*, SG24-5474
- ▶ *IBM Enterprise Storage Server*, SG24-5465

Other resources

These publications are also relevant as further information sources:

- ▶ *Building Storage Networks*, ISBN 0072120509

These IBM publications are also relevant as further information sources:

- ▶ *ESS Web Interface User's Guide for ESS Specialist and ESS Copy Services*, SC26-7346
- ▶ *IBM Storage Area Network Data Gateway Installation and User's Guide*, SC26-7304
- ▶ *IBM Enterprise Storage Server Configuration Planner*, SC26-7353
- ▶ *IBM Enterprise Storage Server Quick Configuration Guide*, SC26-7354
- ▶ *IBM SAN Fibre Channel Managed Hub 3534 Service Guide*, SY27-7616
- ▶ *IBM Enterprise Storage Server Introduction and Planning Guide*, 2105 Models E10, E20, F10 and F20, GC26-7294
- ▶ *IBM Enterprise Storage Server User's Guide*, 2105 Models E10, E20, F10 and F20, SC26-7295
- ▶ *IBM Enterprise Storage Server Host Systems Attachment Guide*, 2105 Models E10, E20, F10 and F20, SC26-7296
- ▶ *IBM Enterprise Storage Server SCSI Command Reference*, 2105 Models E10, E20, F10 and F20, SC26-7297
- ▶ *IBM Enterprise Storage Server System/390 Command Reference*, 2105 Models E10, E20, F10 and F20, SC26-7298
- ▶ *IBM Storage Solutions Safety Notices*, GC26-7229
- ▶ *PCI Adapter Placement Reference*, SA38-0583
- ▶ *Translated External Devices/Safety Information*, SA26-7003
- ▶ *Electrical Safety for IBM Customer Engineers*, S229-8124
- ▶ *SLIC Router Installation and Users Guide*, 310-605759
- ▶ *SLIC Manager Installation and User Guide*, 310-605807

Referenced Web sites

These Web sites are also relevant as further information sources:

- ▶ IBM TotalStorage hardware, software and solutions:
<http://www.storage.ibm.com>
- ▶ IBM TotalStorage Storage Networking:
<http://www.storage.ibm.com/snetwork/index.html>
- ▶ Cisco:
<http://www.cisco.com>

- ▶ QLogic:
<http://www.qlogic.com>
- ▶ Emulex:
<http://www.emulex.com>
- ▶ Finisar:
<http://www.finisar.co>
- ▶ Veritas:
<http://www.veritas.co>
- ▶ Vixel:
<http://www.vixel.com>
- ▶ Tivoli:
<http://www.tivoli.co>
- ▶ JNI:
<http://www.Jni.com>
- ▶ IEEE:
<http://www.ieee.org>
- ▶ Storage Networking Industry Association:
<http://www.snia.org>
- ▶ Fibre Channel Industry Association:
<http://www.fibrechannel.com>
- ▶ SCSI Trade Association:
<http://www.scsita.org>
- ▶ Internet Engineering Task Force:
<http://www.ietf.org>
- ▶ American National Standards Institute:
<http://www.ansi.org>
- ▶ Technical Committee T10:
<http://www.t10.org>
- ▶ Technical Committee T11:
<http://www.t11.org>
- ▶ eServer xSeries 430 and NUMA-Q Information Center:
<http://webdocs.numa.q.ibm.com>

How to get IBM Redbooks

You can order hardcopy Redbooks, as well as view, download, or search for Redbooks at the following Web site:

ibm.com/redbooks

You can also download additional materials (code samples or diskette/CD-ROM images) from that site.

IBM Redbooks collections

Redbooks are also available on CD-ROMs. Click the CD-ROMs button on the Redbooks Web site for information about all the CD-ROMs offered, as well as updates and formats.

Index

Numerics

1394b 23
1x9 Transceivers 21, 27
2062-D01 187
2062-D07 186
2062-T07 187
24-bit addressing 42
24-bit port address 37
6227 149

A

AAA 238
ACC 44
access control 232
access fairness mechanism 33
active connections 273
adapter cable 28
Adapter Hard Loop ID 179
adapter timer settings 160
Address Resolution Protocol 218
addressing scheme 42
AFS 114
aggregate 210
aggregate bandwidth 199, 208
air temperature 204
AL_PA 36, 44
 priority 37
aliases 58
analyze switch fabric configuration 244
analyzing end-to-end connectivity 243
ANSI 19, 75
API 74, 215
application availability 272
application programming interface 215
Application Programming Interfaces 74
application-specific integrated circuit 20
arbiter 204
arbitrated loop 30, 97
Arbitrated Loop Physical Address 36
arbitration 32, 36
arbitration protocol 32
area 43
ARP 218

AS/400 115
ASIC 20
audit trail 134
authentication, authorization, and accounting 238
auto mode 190
automatic failover 270
auto-sensing 186, 188
autosensing 207
availability 270
available addresses 44
AWC 163

B

backplane 20, 193, 203
balancing 92
bandwidth 40, 108
bandwidth utilization 210, 301
BB_Credits 261
benchmark 73
binding 154
bit error rate 261
blade architecture 307
blocking 87
boot function 163
bottleneck 277, 282
bound 154
bridging 29, 329
broadcast 40, 43
buffer credit 152
buffer credits 199, 201, 209, 304
buffer memory 151
building blocks 269
bus arbitration 97
business continuance 270
business continuity 7
business recovery 6

C

C-6460 166
cabinet 134
cabinet key 134
cabinet protection 144
cable identification tag 140

- cable management 134
- cable protection 144
- cable routes 305
- cable routing 125
- cable supports 134
- cable tag naming standard 140
- cable ties 134
- cable types 125
- cascading 61, 92
- cascading switches 277
- Centaur ASIC 153
- Centaur ASICs 153
- central arbiter 195
- centralized management 17
- change record 140
- chassis 186, 202
- Cisco 185
- Cisco Fabric Manager 215
- Cisco IOS CLI 215
- Cisco SAN capabilities 210
- cladding 127
- CLI 226, 239
- clock module 193
- clock modules 203
- closed-ring topology 40
- closing a loop circuit 39
- cluster 15
- clustering software 272
- clustering solution 303
- clustering solutions 111
- coating 127
- color coding schemes 125
- common interface model 65
- CompactFlash 196–197, 204
- complex design 269
- complex switched fabric 49
- concepts 85
- configure hardware components 226
- configure ports 226
- congestion 213
- congestion control 213
- connection authentication 120
- Connection Options setting 181
- connectivity 277
- connectivity problem 258
- core 127
- core switches 280
- core-edge design 280
- coupling switches 88

- crossbar switches 199
- crossbar switching fabric 196
- cut-through logic 48

D

- d balancing 273
- dark 134
- dark fiber 130
- data communications fiber 125
- data migration 16
- Data Path Optimizer 112
- data sharing 12
- data transfer rate 90
- database synchronization 92
- DB-9 24
- decrypt 119
- dedicated fibers 100
- defacto standards 75
- default VSAN 236
- default zone 231
- degraded link 261
- degraded performance 91
- delay factor 40
- destination ID 48
- device driver 155
- Device View 216
- diagnostic 243
- dial home 135
- digital certificate 119
- director class 191
- disaster planning 3
- disaster recovery 4
- disaster tolerance 16, 270
- disciplines 123
- discovering devices 223
- discovery process 223
- distance 209
- distance limitations 124
- distance solution 300
- distance solutions 100
- distributed SAN 300
- distributing traffic 210
- documentation 141
- domain 43
- domain number routing 48
- downstream 34
- downtime 4
- DPO 112

- Dragonfly ASIC 152
- driver registry settings 159
- Driver Trace Mask 159
- dual fabrics 291
- dual fabrics over distance 295
- dual redundant supervisor modules 186
- dual SAN fabrics 292
- dual supervisors 196
- DWC 163
- DWDM 102

E

- E_D_TOV 62, 160
- E_Port 20, 29, 49, 88, 96, 188, 200, 208, 211
- education 145
- EISL 189, 211
- elxcfg 155
- elxcfg.exe 155
- Emulex 149, 151
- Emulex Superfly 151
- encrypt 119
- encrypted 239
- encrypted tunnel 121
- encrypting 222
- encryption 119, 302
- end-of-frame 37
- EOF 37
- ERP 4
- error detect time-out value 62
- ESCON 113, 131, 304
- Ethernet 216
- ethernet cables 133
- Execution Throttle 180
- expandable fabric 88
- Extended Control Block 180
- Extended Error Logging 180
- Extended ISL 189
- Extended Remote Copy 110
- extenders 101
- external boot function 163
- EZ Fibre 167

F

- F_BSY 62
- F_Port 20, 29, 189, 200, 208
- F_Ports 43
- fabric design 276
- Fabric login 44

- fabric login 44
- fabric port layout plan 140
- Fabric pWWN 231
- fabric services 233
- Fabric Shortest Path First 50
- fabric topology 224
- Fabric View 216
- failover 191, 271
- failures 258
- fairness algorithm 36
- fan assembly 195, 204
- fan-out 281
- FAST!UTIL 179
- FasT-MSJ 178
- fault tolerance 272
- FC cabling 19
- FC64-1063-N 164–165
- FC-AL 32, 97
- FCC 213
- FCE-1473 165
- FCE2-6412 165
- FCE-6410 165
- FCI-1063 164
- FCIA 77
- FCIP 198, 206
- FCIP Activation 198
- FC-LE 97
- FC-PH 105
- FC-SW 41, 96
- FC-SW2 96
- fiber optic interconnects 20
- fiber-optic transceivers 22
- Fibre Alliance 75
- fibres cable routes 305
- Fibre Channel
 - point-to-point 31
- Fibre Channel analyzer 190, 263
- Fibre Channel analyzers 249
- Fibre Channel Arbitrated Loop 32
- Fibre Channel cabling 19
- Fibre Channel diagnostic tools 262
- Fibre Channel Industry Association 75
- Fibre Channel Shortest Path First 50
- Fibre Channel Standards 85
- Fibre Channel Switched Fabric 41
- Fibre Channel topologies 30
- FICON 131
- field-replaceable modules 24
- Firewalls 121

- firmware 163
- FL_Port 20, 29, 35, 189, 200, 208
- FL_Ports 43
- FlashCopy 309
- flexible fabric switch 188
- floating WWN 155
- FLOGI 44, 46
- floor plan 124
- flow control 213
- flow control mechanism 215
- footprint 205
- Forward Congestion Control 213
- frame header 42
- frame holdtime 62
- Frame Size setting 179
- frame transmission 39
- free port addresses 45
- FSPF 50–51, 235
- full duplex protocol 31
- full mesh 278
- fusion 129
- Fx_Port 190

G

- G_Port 29
- G_Ports 88
- GBIC 21, 23
- general concepts 269
- geographically dispersed cluster 111
- Gigabaud Link Modules 26
- Gigabit Ethernet 23
- Gigabit Interface Converters 21, 23
- Gigabit Link Modules 21, 26
- gigabit transport 20
- GLM 21, 26
- graphical representation 228
- gridlock 282
- groups 239

H

- hackers 6
- Hard Loop ID 179
- hardware enforced 233
- hardware zoning 57
- HBA 149
- HBA settings 161
- HBAnyware 157
- heat damage 136

- high availability 211
- high priority status 213
- highly available SAN 303
- high-speed switching 42
- holdtime 62
- hop count cost 92
- Hoppy 62
- hops 61–62
- Host Adapter BIOS 179
- Host Bus Adapters 149
- host bus adapters 270
- hot-swappable fan 186
- HSSDC 24
- hub cascading 34
- human error 138

I

- i 201
- I/O circuits 20
- I/O Priority Queuing 16
- identity 239
- IEEE 42, 53
- IETF 75
- image pair 38
- implicit transfer 236
- in order frame delivery 94–95
- increased availability 23
- intangible damages 6
- interconnection topologies 30
- interface module 205
- internal visibility 256
- interoperability 73
- interoperability matrix 252
- interoperability mode 241–242
- Interrupt Delay Timer 181
- intersite cable route 133
- inter-switch links 61
- investment protection 185
- IN-VSN 304
- IP Line Card 198, 206
- IP protocol 177
- IP services 212
- IPSec 121
- iSCSI 198, 206
- iSeries 115
- ISL 61, 211
- ISL segmented 88
- ISL synchronization process 88

ISL traffic 298
islands of information 17
ISO 75
isolated VSAN 236
ISP 6

J

JFS 114
Jiro 75
JNI 164
JNI EZ Fibre 167
jumper cable 128, 130
just-in-time storage 66

K

kangaroo 62
key access 134
keys 119

L

L_Port 29
labelling 140
LAN 73
LAN free data movement 14
lasers 24
latency 32, 107, 133, 302
LC-SC 28
LED status 159
legacy tape software 154
light 125
light loss 129
Light Pulse LP7000E 151
limited addressing 97
limited distance 97
link cost 91
link lengths 104
link state change 92
LIP 33, 44–45, 180, 313
LIP Full Login 180
LIP Reset 180
LIP's 33
LISM 34
load balance 210
logic control 186
logic functions 20
logical consolidation 11
logical start topology 31

Logical Unit Number 60
Login Retry Count 180
long wavelength 22
longer distances 100
longwave 125
loop
 private 35
 public 35
loop addressing 37
loop circuit 39
loop device 312
loop initialization 33
Loop Initialization Master 34
Loop Initialization Primitive 33
loop initialization process 180
loop protocols 33
Loop Reset Delay 179
loop solution 312
loop tenancy 36
loopback 166
loops 35
loop-specific protocols 33
loss of synchronization 105
low bandwidth LAN 98
low-loss 25
lp6dutil 156, 163
LP7000E 151
LP8000 152
LP8000S 152
LP9002DC 153
LP9002L 153
LP9002S 153
lputil 156
LUN 60, 180
LUN level masking 167
LUN mapping 154
LUN masking 60, 304, 314
LUN-Level Zoning 167

M

Magstar 16
MAN 133, 137
managed service providers 301
management traffic 222
masking 60
master 34
maxphys 172–173
MDS 9000 xix, 185

- MDS 9216 186
- MDS 9216 Multilayer Switch 187
- MDS 9509 186
- MDS 9509 Multilayer Director 186
- mechanical splices 129
- Media Interface Adapters 21, 27
- members 58
- memory 58
- memory arrays 20
- meshed network topology 278
- Metropolitan Area Network 133
- MIA 21, 27
- MIB 69
- Mini Port Driver 155
- miniport driver 151
- mirroring 109, 273
- mixed fiber diameters 130
- modular chassis 200, 208
- monitor 209
- monitor bidirectional traffic 251
- monitoring network traffic 245
- multicast 95
- multilayer network 185
- multimedia 98
- multi-mode 125
- Multiple Allegiance 16
- multiple paths 270–271
- MultiPulse 157

N

- N_Port 29, 37, 44
- N_Ports 20
- name server 60
- name server table 54
- naming conventions 138
- naming standards 134, 138
- native FICON 113
- network fabric 222
- network traffic 190
- NL_Port 29, 37, 44
- non-blocking device 306
- non-broadcast 40
- non-intrusive 190
- non-structured cables 128
- non-trunking ports 236

O

- off-campus 130

- Open Standards Fabric Initiative 75
- Open Systems Adapters 131
- open-init 33
- operating system logs 158
- optical connectors 25
- ordered sets 33
- OSA 131
- OSFI 75
- oversubscribing ISLs 92
- oversubscription 281–282

P

- pair of E_Ports 88
- Parallel Access Volumes 16
- Parallel Sysplex 114
- parity algorithm 109
- partial mesh 278
- passwords 143, 219
- patch panels 129
- Path and dual-redundant HBA 271
- path failover 271
- pathway 125
- PAV 16
- Peer to Peer Remote Copy 304
- Peer-to-Peer Remote Copy 16
- performance problems 260
- persistent binding 154
- phantom 47
- phantom address 47
- phone line 135
- phone line bottlenecks 135
- phone sockets 135
- physical consolidation 10
- physical layer problems 158
- physical location 123
- physical specifications 202
- pin through hole 21
- platform independence 63
- PLOGI 38, 46–47
- point-to-point 30–31
- pooling 66
- port 43
- port address 42–43
- port density 199
- Port Down Retry Count 180
- Port Driver 155
- port driver logs 158
- port driver utilities 155

- port formula 282
- port layout plan 140
- port login 38
- port modes 88
- port module 192
- Port VSAN membership 237
- Port World Wide Node Name 231
- port-based zoning 58
- Port-Channeling 298, 302
- Port-Channels 191, 210, 226
- positional map 33
- power consumption 194
- power down 142
- power on Sequence 142
- power outlets 134
- power supplies 136, 194, 203–204
- power supply options 194
- PPRC 304
- pre-planning activity 123
- prevent unauthorized access 309
- priority 37, 213
- Private Fibre Channel devices 45
- private host 48
- private key 119
- private loop 35, 37
- PRLI 38
- process login 38
- propagation delay 40
- proprietary management solutions 63
- protecting data 308
- PTH 21
- public arbitrated loop 189
- public key 119
- public loop 35, 37
- push-pull 25
- pWWN 231

Q

- QLA2100 176
- QLA2100F 176
- QLA2200 177
- QLA2200F 176
- QLA2300 178
- QLogic 176
- QoS 213
- quality of service 213
- quench control 214
- quench message 214

- queues 213
- queuing logic 20

R

- R_A_TOV 62, 160
- R_RDY 106, 261
- RADIUS 121, 219, 238, 241, 309
- RAID 180
- RAID levels 109
- Receiver Ready 261
- Redbooks Web site 334
 - Contact us xxii
- redundancy 270
- redundant fans 192, 203
- registered devices 52
- Remote Authentication Dial-In User Service 121
- repeaters 100–101
- Request Product Quotation 150
- resilience 210
- resource allocation time-out value 62
- resource management 66
- RIO Operation Mode 180
- rodent control measures 144
- role-based authorization 240
- roles 239
- routing algorithms 92
- routing decision 48
- routing logic 42, 48
- RPQ 150
- running disparity 37

S

- SAN disciplines 123
- SAN fabric cabinets 124
- SAN implementation 270
- SAN inventory 124
- SAN islands 186, 304
- SAN management 63
- SAN platforms 113
- SAN software management 63
- SAN standards 73
- SANblade 178
- SANblade Control FX 178
- SANergy 13
- SANMark 176
- SANmark 77
- SANmark Qualified Program 77
- SANsurferTool Kit 178

- SBUS 168
- SC connectors 24
- SC type 25
- scalable solution 192
- SCSI limitations 97
- SCSI Trade Association 75
- SCSITA 75
- sd_max_throttle 172
- SD_Port 190, 201, 209, 249, 251
- SDD 112, 271, 304
- secondary supervisor 196
- secure fabric 120
- secure machine room 145
- Secure Shell 121, 238
- Secure Sockets Layer 121
- security 56, 60, 134, 142, 226, 308
- security breaches 115
- security layers 116
- security solution 308
- seed switch 223
- segmentation 55
- segregation 233
- sending frames 107
- serial optical converters 21
- server clustering 15
- server free data movement 14
- servers 151
- setup routine 218
- SFF 21
- SFP transceivers 201, 209
- shared bus 32
- short wavelength 22
- shortwave 125
- signal travel times 104
- signature 56
- silica glass 127
- simple fabric 269
- Simple Network Management Protocol 121
- single-level storage 115
- single-mode 125
- slots 186
- small form factor transceivers 21
- SNIA 75–76
- SNMP 223, 226, 239
- snooper port 209
- SNS 44–46
- soft zoning 142
- software zoning 58
- Solaris driver 166

- source interface types 247
- space management 66
- SPAN 245, 251
- SPAN destination port 190
- SPAN session 248
- SPAN source 245
- spare ports 298
- SSH 121, 238
- SSL 121
- stable monitoring mode 33
- storage intelligence 185
- Storage Network Industry Association 75
- Storage Networking Industry Association 76
- StorWatch 17
- striping 109
- structured cables 128
- structured cabling system 130
- Subsystem Device Driver 112, 271, 304
- Summary View 216
- supervisor exchanges 196
- supervisor function 187
- supervisor module 186, 191, 195, 203–204, 216, 220
- supervisor modules 192, 228
- supported HBAs 150
- switch authentication 120
- switch cascading 61
- switch frame buffering 48
- switch health option 243
- switch management 238
- switch maximum
 - 239 switches in fabric 61
 - 7 hops allowed 61
- switch port addresses 42
- switch port analyzer 190, 245
- switch port capability 88
- switched fabric 30, 41
- switching module 197, 204, 207
- switching modules 196, 205, 228, 306
- symbolic names 58
- system outage 5
- System/38 115

T

- T_Ports 304
- T11 97
- tape pooling 13
- Target Reset 180

- TE_Port 189, 200, 208, 211
- TE_Ports 294, 304
- Telnet 239
- temporary loop master 34
- terminology 85
- tier-layer design 278
- time-critical traffic 213
- Tivoli 14
- Tivoli SAN Manager 17
- Tivoli Storage Manager 17
- Tivoli Storage Network Manager 17
- TL_Port 29, 189, 200, 208, 313
- TLS 121
- topology changes 51
- topology database 92
- trace tools 263
- traffic patterns 91
- Transceiver 201, 209
- transceivers 21, 27
- translation table 47
- translative loop port 208
- translative mode 48
- transparent failover 273
- Transport Layer Security 121
- transport protocols 188
- Tri-Rate 201, 209
- Tri-Rate SFP 209
- troubleshoot fabric connectivity 243
- troubleshooting features 243
- troubleshooting the SBUS HBA 169
- troubleshooting tips 255
- troubleshooting tools 244
- trunk cables 128
- trunking 211
- trunking E_Port 189
- tunneling 198, 206

U

- U_Port 29
- ultraviolet 127
- unauthorized access 143
- unidirectional 128
- UNIX Servers 114
- update mechanism 92
- upstream 34
- UV 127

V

- velcro strips 134
- vendor identifiers 53
- Virtual SAN xix, 185, 233, 235, 292
- virtual SAN 222
- Virtual Tape Server 16
- VSAN xix, 185, 211, 214, 222, 225, 233, 247, 292, 298, 309
- VSAN attributes 237
- VSAN manager 238
- VSAN membership 237
- VSAN trunking 189
- VSAN-based filtering 248
- VSANs 200, 211
- VTs 16
- vulnerabilities 116

W

- WAN 73
- warning messages 204
- World Wide Name 42, 45, 157, 168
- world wide name 52
- World Wide Node Name 44
- World Wide Port Name 44
- world wide port name 53
- WWN 42, 44, 52, 157, 168, 231
- WWN address 42
- WWN zoning 58
- WWNN 154
- WWPN 44, 52–53, 154

X

- X3T11 19
- XRC 110

Z

- zone 55
- zone naming standards 141
- zones 224
- zoning 229, 231
- zoning restrictions 232



IBM SAN Survival Guide Featuring the Cisco Portfolio

(0.5" spine)
0.475" <-> 0.875"
250 <-> 459 pages



Redbooks

IBM SAN Survival Guide

Featuring the Cisco Portfolio

**Protect your data
with an IBM SAN**

**Build a SAN too
tough to die**

Survive and conquer

As we all know, large ocean going ships never collide with icebergs. However, occasionally life deals out some unexpected pleasures for us to cope with. Surviving any disaster in life is usually a lot easier if you have prepared adequately by taking into account the likely problems, solutions, and their implementation.

In this IBM Redbook, we limit ourselves to those situations in which it is likely that a SAN will be deployed. We present the IBM SAN portfolio of Cisco products, going a little under the surface to show the fault tolerant features that they utilize, and then describe solutions with all these features taken into account. Each of these solutions was built on practical experience, in some cases with cost in mind, in some cases with no cost in mind. Any well thought out SAN design will have taken each of these concerns into account, and either formulated a solution for it, or ignored it, but nonetheless understanding the potential exposure.

With these points in mind, in this redbook we have two objectives: to position the IBM SAN products that are currently in our portfolio; and to show how those products can be configured together to build a SAN that not only allows you to survive most forms of disaster, but also provides performance benefits. So, make sure that you know what to do if you hit an iceberg!

**INTERNATIONAL
TECHNICAL
SUPPORT
ORGANIZATION**

**BUILDING TECHNICAL
INFORMATION BASED ON
PRACTICAL EXPERIENCE**

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks